TESTING SECOND LANGUAGE WRITING IN ACADEMIC SETTINGS

Elizabeth Margaret Hamp-Lyons

Doctor of Philosophy
University of Edinburgh
1987

DECLARATION

I declare that this dissertation is entirely my own work

Signed

ACKNOWLEDGEMENTS

Many people have helped me in the process that made this product a reality. My dissertation committee, Alan Davies (Chair), Clive Criper and Ron Asher each made invaluable and unique contributions, as did Alan Davies and Clive Criper as Directors of the Edinburgh ELTS Validation Project, and Clive Criper as Director of the Institute for Applied Language Studies, University of Edinburgh. Debby Keller-Cohen, Director of the English Composition Board, University of Michigan, gave great support in the final year.

Thanks to all my friends and colleagues at the British Council and the University of Cambridge Local Examinations Syndicate, Ian Seaton, Clive Bruton, Gill Westaway, John Foulkes, Trisha Aspinall, Alan Moller, Peter Hargreaves and Terry Toney; to my raters Graham Cawood, Lindsay Criper, David Crosby, Mairin Dormer, Mike Lyons, Elizabeth McLelland, and to those on the MSc in Applied Linguistics/ELT in 1985-86; to Tony Lynch for data collection; to Myint Su, Lesley Shields and Margaret Love in English Language Testing Research; to the testing brown baggers, Rosemary Baker, Gibson Ferguson and Robert Hill.

To Charles Alderson, Ben Heasley, John Maher, Ian McGrath and Cyril Weir as personal friends as well as wonderful professional colleagues, special thanks.

I have benefitted tremendously from comments by Charles Alderson, Kathi Bailey, Hans Dechert, Grant Henning, Robert Kaplan, Mike Lyons, John Maher, Greg Myers, Alastair Pollitt, Albert Pilliner, Pauline Robinson, Larry Selinker, John Swales, Elaine Tarone, Cyril Weir and Vivian Zamel.

To my parents, who know why;

To my little son, Chris, who has never known a mother with a free weekend; to my big son Nick for coffee and babysitting;

Most of all, to Mike, for everything - and then some: THANK YOU.

ABSTRACT

While the direct assessment of proficiency in writing through the collection and evaluation of one or more writing samples is a common activity in educational systems, and has been extensively and intensively researched, there is still much to be learned. The context of this study is the evaluation of the writing proficiency of overseas, mainly postgraduate, applicants to British tertiary education institutions. The empirical study investigated two competing claims:

(1) that an appropriate writing test in this context should be specifically related to the content of the testee's academic discipline;
(2) that an appropriate writing test in this context should require expository writing of a kind generally encountered in the academic community across disciplines.

Following investigation of the constructs of 'writing', 'proficiency' and 'specific purposes' and the establishment of formal criteria from language testing, it was hypothesized that:

(1) Scores assigned to the writing of non-native postgraduates at British universities when writing on discipline specific (SAP) topics would not be significantly different from the same subjects' scores when writing on general academic (GAP) topics.
(2) Scores assigned to the same subjects for two 'parallel' SAP questions would share more variance than scores assigned to these subjects for one SAP and one GAP question.
(3) Single-rater scores resulting from the operational scoring procedure would not be adequately reliable.
(4) A three-rater aggregate score would be adequately reliable.

The writing of 111 subjects in five 'Modules' on two SAP questions and one GAP question was studied; it was found that operational scores were unreliable but aggregate scores were adequately reliable. Results of the study presented a conflicting pattern, with significantly higher mean scores for SAP than GAP in some cases, but with more significant correlations between SAP and GAP than between SAP and SAP. Rarely was more than 60% of score variance accounted for in any interaction. It was suggested that a consistent SAP ≠ GAP distinction is not being maintained in the test design.

The key writing test variables of scoring procedure, reader variables, essay test task design and writer variables were intensively studied in an attempt to move toward a better understanding of what was causing the inconsistent results. It was found that the scoring procedure used general rather than specific academic criteria, and that raters were applying these criteria to SAP as well as GAP writing tests. Close study of raters gave no indication that they were recognising and valuing SAP responses from writers. Issues of task design and task difficulty were approached through the study of writers' responses, and some progress was made in understanding the characteristics of an SAP response from a writer. The three writing tests exhibited no clear SAP/GAP distinction, and neither difficulty levels nor task demands exhibited uniformity.

It is suggested that until the scoring procedure and criteria are made more valid, raters are trained to make SAP judgements, and tasks which are more validly SAP are designed, no firm conclusion as to which of the competing models is more valid can be drawn. Until that time, there appears to be little support for the use of purportedly specific academic purpose rather than general academic purpose writing test tasks.

CONTENTS

CONTENTS

CONTENTS

CONTENTS

CONTENTS

CONTENTS

CONTENTS

PREFACE

Background and rationale

This study is set in the context of a larger study (Edinburgh ELTS Validation Project: Criper & Davies, 1981 - 1986) of the English Language Testing Service, the testing service provided by the British Council. The British Council is Britain's principal funding body for scholarships for overseas applicants to education institutions and courses, and the English Language Testing Service (ELTS) as originally introduced and operated was a test for applicants to academic tertiary educations (universities, polytechnics, and advanced professional courses). Because of the undergraduate entrance requirements of British universities, almost all applicants are postgraduates or equivalent. [1]

The ELTS is a two-tier test. In the first tier there are two multiple-choice tests, one of which is a reading test and the other of which is a listening test. These two tests are taken by every ELTS testee. In the second tier there are three tests: a multiple-choice study skills test, a direct writing test; and an oral interview. In this second tier there are six choices or 'Modules', and every testee must choose one of these. The six Modules are Life Sciences, Medicine, Physical Sciences, Social Studies, Technology and General Academic. The first five of these were designed to conform to the five largest groups of applicants for British Council scholarships, while the last, General Academic, was intended for all those applicants who did not fit easily into the other Modules.

The ELTS was designed to put into practice three theoretical positions or constructs of how language proficiency is composed. The first is relatively uncontentious: the ELTS views language proficiency as divisible on the skills dimension, and it has separate tests of reading, listening, writing, and speaking. The ELTS takes this construct further than other

tests, certainly further than any other test at the time of its introduction.

The second construct underlying the ELTS is a view of language proficiency as divisible into 'general' and 'study' proficiency. The second tier of the test is the test of study proficiencies: M1, 'Study Skills', is intended to test reading and other text skills with a focus on those 'micro-skills' amd 'micro-functions' which are used in academic study contexts; M2 'Writing' is intended to test writing in academic settings, and M3 'Oral Interview' is intended to test speaking skills in academic settings.

Third, and perhaps most contentiously, the ELTS divides language proficiency on a subject, or discipline, dimension, through the six Modules referred to above. Thus in the ELTS, there is a construct which assumes that there is such a thing as Life Sciences for study purposes, which is distinct from Medicine for study purposes, etc.

While a test such as the ELTS offers rich opportunities for research, even the most cursory examination of M2, the writing test, showed that there were particularly strong imperatives for detailed research into it, and for generalizing from the results of such research to the testing of writing in similar contexts, i.e., other contexts where testing writing in academic settings involved a choice between writing tests for general academic purposes and writing tests for specific academic purposes.

Before an empirical investigation could proceed, however, the problem had to be put into context through the exploration of a number of preliminary concepts and constructs, and some hypotheses had to be determined.

Chapter 1

Because the written form is so often used as a means of testing knowledge in content areas, it may be forgotten that the medium is itself an area of knowledge and skill. When a person is required to write so that her mastery of a subject matter can be assessed, or so that generalizations can be made from her performance to her readiness for study of a particular kind or at a particular level, it is not simply a set of re-ponses which can conveniently be measured against a criterion, a norm or other measurement scale which are collected. In addition, an authentic and personal response is elicited Writing is something real, s mething people actually do, it is not a contrived response-type existing only in examination halls In investigating any writing test, then, serious attention must be paid to an understanding of the activity of writing, or composing· what characterizes it? how we can recognise successful outcomes of it? Writing performance cannot be measured until writing is understood as a construct, and this is the focus of the first part of Chapter 1.

Writing as e sment ha always been problematic, and there is no reason to supp e that the testing of the writing in English of non-native Engli h speaker presents fewer problems than the testing of the writing of native spe kers of the language. An understanding of the fundamental oncepts of language testing - reliability, validity, practicality, and backwa h - provides a sound starting point from which it is possible to inve ti ate practices and problems in the testing of writing in the context of this study, and this is the focus of the second part of Chapter 1.

In the empirical investigation it will be necessary to keep in mind the principles about the composing process as far as it has been possible to derive these from Chapter 1. The following is an attempt to derive these principles:

1. *Writing is a heuristic procedure· through writing learners learn not only to write, but to think and feel (Murray,1978; Zamel, 1983);*

2. *Writing is not either 'good' or 'bad': acquiring skill in writing is a developmental process, and it is possible to judge what stage of development a writer is at by exploration of such factors as distan e from the self (Wilkinson 1978 a+b), complexity of T-units (Hunt, 1965, etc) and number of error-free T-units (Stewart, 1978);*

3 *Writing is interactive: a writer needs a reader, and the (perceived) nature of the reader influences the writing pro ess and product (Flower, 1979; Young et al 197);*

4. *Writing is a multi-dimensional activity: a writer goes through a number of stages (Weaver, 1978; Britton et al, 1978) and the stages are not (necessarily) linear (Flower and Hayes, 1981; de Beaugrande, 1983). In addition, a number of processes can occur during each stage. It is also purposive: purposeless writing, or writing initiated by someone other than the writer where the writer does not accept or 'value' the purpose, will suffer in quality (Weaver, 1978);*

5. *Writing is normative: throughout the writing process of the good writer runs an awareness of the need to conform to a range of norms (Flower, 1979; Nystrand, 1983 (b); Shaughnessy, 1977). This principle explains the continuing emphasis on the product rath r than the process in the teaching and evaluation of writing· product norms can be specified and quantified in relatively precise ways, unlike process-related principles;*

6. *All writing beyond handwriting practice is also composing· the writer must express something of herself in every composition.*

In Chapter 2 it will also be necessary to keep in mind the principles of language testing which are summarised here:

1. *Language tests can be of different types and for different purposes. the purposes for which testing is being carried out will determine the type of test and the rigorousness with which other testing principles are applied,*

2. *Unless a language test is reliable its other characteristi-s cannot be meaningfully investigated. The level of reliability which is demanded of a test depends on its type and the purpose of testing,*

3 *Reliability is a necessary but not sufficient hara teristic of a language test. validity is the central haracteristic, but we cannot measure it until we have achieved reliability,*

4. *Con truct validity is the most difficult validity to stablish for a language test, yet it is the validity langu ge tests need most of all.*

Chapter 2

The use of direct writing samples as the basis for judgements has been ommon in Britain and the United States of America since the Victorian era, and has been problematic throughout that period. Direct writing as essment is typically claimed to have inherent validity; problems are usually seen as relating to reliability Chapter 2 looks in detail at the reliability and validity of direct writing assessment as well as at its practicality and ba kwash.

In the terms of this study, an essay test is the collection of a direct writing sample or samples from candidate writers in a highly structured test context, in response to an assigned task, which sample(s) is/are then read and rated according to a more or less precisely specified procedure by one or more qualified judges. In this definition, we have four key variables contributing to what, for the sake of convenience but at the ri k of accuracy, is referred to throughout as the 'essay': the task; the writer; the scoring procedure; the rater (the terms 'rater' and

'reader' are used interchangeably throughout this study). These are dealt with in reverse order in Chapter 2.

A complex of factors which combine and interact to make a particular rater on a particular occasion respond in ways which are not only more or less reliable but also, in any writing assessment which is purposive, more or le valid, are discussed. A large range of scoring procedures is rveyed and it is suggested that a choice of scoring procedure will al o have effects not only on the reliability attributable to the asses ment, but also on its validity.

Th re is not y t a research methodology for the study of writer variables or their impact, nor even a developed categorisation of writer variables as di tinct from task variables. There is at present no certain way to determine the amount of true variance in a writer's essay test score whi h is due to theoreti ally predictable characteristics of the writer, a oppo ed to the interaction between the writer, her performance on the specific occasion, and a range of other characteristics, notably ch racteristics of the t st ta k. Investigations of the ways that writers' responses are shaped by the tasks on writing test are in their infancy, but the early research suggests a number of important design fa tors which can be applied in shaping a writing test task to give every testee an equal opportunity to give her best performance.

Clearly, the principles for the testing of writing remain the same for the testing of the writing of second language writers. There is little research evidence to indicate whether or not findings from studies of English L1 writers and their writing on essay tests can be applied to ESL writers writing essay tests in English. Throughout Chapter 2, however, insights from first language studies and from second/foreign studies are interwoven, and it is suggested that not only are there no contradictions in the issues, methods or results reported, but further,

this interweaving provides a much fuller picture than would be possible if only the research from one or the other were surveyed.

Chapter 3

The concern of this study is not simply the empirical investigation of a direct writing assessment, but of a direct writing assessment designed for a very specific context and based on strong and controversial claims, as was suggested in the first section of this Preface and as is detailed in Chapter 4. The test battery as a whole rests on a claim that to test 'general language proficiency' is insufficient for the context in which this test operates. In this context, it is claimed, language proficiency should be tested for each skill area separately; for study purposes rather than general purposes; and, in the second tier of the test, for specific academic purposes rather than for general academic purposes, i.e., through discipline-specific test materials. The bases for claims of general language proficiency (UCH, or the Unitary Competence Hypothesis) and of divisible language proficiency (DCH, or Divisible Competence Hypothesis) are examined in the first part of Chapter 3, and the arguments for a direct writing assessment which is intended to be not only an **academic writing** test but further, a **specific academic writing** test, are placed in the context of the debate of the past ten years over the construct of language proficiency.

The second half of the Chapter seeks to define and classify 'specific purposes' (ESP, English for Specific Purposes) and 'academic purposes' (EAP, English for Academic Purposes) teaching and testing, firstly in general, and secondly with the focus of attention narrowed to an attempt at an understanding of what might be meant by 'academic purpose writing' and 'specific academic purpose writing'. The advantages claimed for 'academic purposes' and 'specific purposes' teaching and testing are that they are more relevant to learners' knowledge and needs, and that because learners of English will be learning the language, and demonstrating

their proficiency in it, through material with which they are already familiar they will be able to perform better. These claims form the basis for the research questions in the main study reported in Chapter 4.

Chapter 4

When the ELTS was being developed, interest in the direct testing of writing was growing in general, and the British Council's commitment to ESP te ting and to communicative testing meant that writing really had to be tested through direct performance. At the same time, however, language testing was undergoing something of a reaction to the psychometric-structuralist p riod and reliability was lower on the agendas of many test d velop rs than validity (in the restricted, non-statistical sense of conforming to a certain view of how the test should be). In the case of M2, the concern of the test developers was primarily with content validity, and questions of score reliability were little considered The result of this was that within two years of the introduction of the ELTS, objections were being raised to the writing test on the grounds of its poor reliability.

Such objections to writing tests are by no means new. A much newer objection, though, was to the concept of discipline-specific tests. The ELTS writing test, M2, like the other two parts of the second tier of the ELTS, is spoken of as 'discipline-specific' and validity is claimed for it on that basis. It has yet to be shown, however, that the ELTS is in fact a discipline-specific test, as opposed to a more or less general academic test using discipline-related texts. It has not been shown in what ways, if any, the performance the test elicits is actually discipline-specific. Indeed, the Edinburgh ELTS Validation Study (Criper and Davies, 1986) in a content validity study of G1, G2 and M1, the three objective components of the test, found little to suggest that the test content, or the kinds of performance elicited, were either discipline-specific or even exclusively academic.

In the case of the writing component, M2 (specifically, M2Q1, since this study does not investigate the second question of M2, which was at that time severely restricted in time for writing and writing task type, and since performance on M2Q2 only affected the final score for M2 by + or - 6%, i e., by a maximum of half a band) it would be necessary to investigate not only the reliability of the test but also its validity

Chapter 4 describes the first part of ELTS M2 (i.e., question 1) and considers the expectations which a specific academic purpose ('SAP') writing test such as ELTS M2Q1 must fulfil, before investigating the two research questions. The main research question asks what the effects on writing test scores are when overseas, non-native, postgraduate students at British universities are asked to write on topics closely related to the content of their own discipline (discipline-specific or 'SAP' topics) compared to a topic accessible to all members of the university community (discipline-free or 'GAP' topics), and how these effects can be accounted for. The subsidiary research question asks whether scores a signed to e say test answers, whether SAP or GAP in nature, by the operation scoring procedure for M2Q1 (1980 - 1985) are adequately reliable for research and operational uses.

The final rationale for the development of a test of the complexity of the ELTS must be that it provides a fairer measure of a testee's language proficiency than the test or tests which it replaces. Thus the rationale for a pecific academic purpose (discipline-specific) writing test such as M2 must be that it yields information which corresponds more closely to what the testee can actually do in regard to the writing required in British postgraduate education than either an indirect test of writing ability or a general academic purpose (discipline non-specific) direct writing test. The information must be at least as reliable as the information yielded by discipline non-specific writing tests, and it should have greater claims to other validities.

The empirical inve tigation indicated that M2Q1 was not satisfactorily reliable as an operational test instrument It found no consistent, significant evidence to show that SAP writing tests advantage testees in the ways predicted by the strong ESP construct,; it also found no evidence that the effects of SAP writing tests as opposed to a GAP writing test fitted a predicted pattern. While SAP writing tests yielded higher scores than a GAP writing test the differences were rarely significant. The two SAP writing tests were not parallel on statistical criteria while in a number of cases GAP had more common variance with one or both of the SAP writing tests than the two SAP writing tests had with ea h other. While all three writing tests were highly correlated in most c es, the amount of variance they shared was not as great as predi ted.

A a result of the empirical investigation it became clear that further investigations were needed, to attempt to discover what variables were operatin to invalidate some of the assumptions made by the ELTS and a cepted for the empirical investigation, i.e., that the tests designated 'SAP' were in fact tests of specific academic purpose writing, and that they were different in important ways from the 'GAP', general academic purpo e, task.

Chapter 5

In the process of carrying out the subsidiary study, of rater/score reliability, it became clear that the ELTS writing test as used at that time could not guarantee adequate reliability for an operational writing te t. There was no possibility of using multiple raters in the British Council ontext: what was urgently needed was a scoring procedure which could provide scores of adequate reliability with only one rater. Investigation of the scoring procedure was also necessary to construct validation. Chapter 3 proposed four key variables for a writing test, of which procedural variables are the first. For a construct valid SAP

writing test we should expect to find evidence that all or several of the e variables are operating along SAP dimensions in the test

An inve tigation of the scoring procedure permits identification of the features of writing it values or does not value, leading to a greater understanding of what is being measured by the test The first part of Chapter 5 d ~cribes the d velopment of the scoring procedure by this researcher and considers what eviden e there is to indicate that M2Q1 has been scored by a SAP pro edure, and also the level of reliability which can be a hieved by a test of writing in academic settings when scored by a single rater in less than ideal circumstances.

The sec nd part f Chapter 5 looks at rater variables through an analysis of t pe re ording of the piloting of the M2 Assessment Guide, and attempts to und rstand how raters respond to a writing test which is described s a SAP writing test Study of the tape recordings permits the rec n tructi n of the criteria which raters use to judge writing ample , and is an important aspect of the search for evidence that M2Q1 is in fact a specific academic purpose test

Chapter 6

Because it was seen in Chapter 3 that task variables operating in any direct writing ssessment are numerous, complex and poorly understood, an unders nding of task variables in this context is approached through study of writers' responses The answers were carefully read a number of times in a search for patterns of response which appeared to be related to components of the task: question, resources and rubric. 'Marked' responses in particular were noted and the concept of the 'challenge' was developed.

Ta k variables and writer variables are interwoven in a search for indications of the ways in which writers respond to tasks and the

qualities which make a task a specific academic writing task rather than a general academic writing task. A tentative system for task analysis and for assessment of task difficulty is developed, and the questions on the three writing tests used in this study are considered on this basis. It cannot, however, be concluded that a satisfactory means of predicting task difficulty on the basis of task analysis in the present state of that art is possible.

Conclusions

The scoring procedure was found to be principally a GAP procedure, raters were found to function as GAP rather than SAP raters, and tasks were found to be poorly designed and without a coherent set of relationships, so that some supposedly SAP tasks displayed more features of GAP tasks and so that difficulty levels varied widely. Under these conditions it is unlikely that findings will be stable enough to permit conclusions about the choice of SAP or GAP writing tests for the testing of the writing in academic settings of overseas postgraduate students to be drawn. It is felt that the variations in score patterns observed in the empirical study are adequately explained by these inconsistencies across task variable , and that the study does not provide evidence to support the existence of an advantage for students tested by a writing test in their disciplinary area over a writing test appropriate to all members of an academic community. However, it is felt that the study has clarified many of the problems of test design for specific academic purpose writing tests

Some suggestions for future research and development are made: in particular a need is identified for research which links studies of disciplinary communities, of writing as process as well as product, and of the development of scoring procedures, drawing on what has been learned in all these areas to inform the development of valid assessment of writing in academic contexts. Research into raters' rating processes

using ethnographic methods such as that in Chapter 5 Section 2, and research into writers' processes using structured interviews and a process parallel to that in the ethnographic study, are identified as being potentially particularly fruitful.

Note

1 Since 1980, when the Edinburgh ELTS Validation Project began, and since 1983, when this study began, the ELTS has undergone a number of changes not described here. Important among these are the development of alternate forms for most components, development of a training package for M3, Oral Interview, and the implementation of an M2 training Manual based on the work reported in Chapter 5 section 1 and on separate development work by this resear her on M2Q2. In that period the Service has expanded to 13,000 candidates annually in 144 test centres in over 90 countries, and an increasing number of undergraduate candidates. Plans are in hand for the development of an 'Undergraduate' Module, for wider publicity and information about the S rvice, and for expansion of the Service to Canada, Australia and the United States of America. In 1987 a new period of research and development of the Service will begin, drawing on the results of studies completed to date.

It should further be noted that, although the expression "the ELTS" has been used in this study, to parallel the expression "the English Language Testing Service", the British Council have established a policy of referring to the test as "ELTS" only. (Gill Westaway, Consultant, ELTS, The British Council· personal communication)

CHAPTER ONE

TWO APPROACHES TO THE PROBLEM: RESEARCH IN WRITING AND COMPOSING; AND IN LANGUAGE TESTING

INTRODUCTION

Because the written form is so often used as a means of testing knowledge of content areas, it may be forgotten that the medium is itself an area of knowledge and skill. When a person is required to write so that her mastery of a subject matter can be assessed, or so that generalizations can be made from her performance to her readiness for study of a particular kind or at a particular level, it is not simply a set of responses which can conveniently be measured against a criterion, a norm or other measurement scale which are collected. In addition, an authentic and personal response is elicited. Writing is something real, something people actually do; it is not a contrived response-type existing only in examination halls. In investigating any writing test, then, serious attention must be paid to an understanding of the activity of writing, or composing. what characterizes it? how we can recognise successful outcomes of it? Writing performance cannot be measured until we have understood writing as a construct, and this is the focus of the first part of this chapter.

In the second part of the chapter, the focus shifts to a survey of some fundamental concepts of language testing which can inform the investigation of the nature and effectiveness of measurements of writing performance in the empirical study. Direct writing assessment has been problematic since it became common in Britain and the United States of America in the mid 1800s. There is no reason to suppose that the direct assessment of the writing of non-native users of English is less problematic: an understanding of the basic expectations of a good

language test provides a sound starting point from which it will be possible to proceed with the investigation of practices and problems in direct writing assessment in later chapters.

1. RESEARCH IN WRITING AND COMPOSING

1.1. Search for a definition

1.1.1. Writing

What is writing? Surely it is not, as Bloomfield (1933) described it, "merely a way of recording language by means of visible marks" (p 31). It is not speech written down. Is it then, as Gabriel Fielding, one of Murray's (1978) writers, described it, "...a voyage, an odyssey, a discovery" (p.101); or another, Lawrence Osgood: "like exploring...as an explorer makes maps of the country he has explored, so a writer's works are maps of the country he has explored" (p.103)? Or does it fit Vachek's more prosaic description: "...a system in its own right, adapted to fulfil its own specific functions which are quite different from the functions proper to a phonetic transcription" (1966: p.157). As Young (1978) suggested, researchers in this field rejected the traditional conception of writing and are exploring a new paradigm (in Kuhn's (1970) sense of a system of widely shared values, beliefs and methods that determine the nature and conduct of the discipline) which requires a new definition.

1.1.2. Composing

We can clarify the question a little by substituting the term 'composing' or 'composition' for 'writing'. The term 'writing' encompasses the recording of any kinds of symbols on paper or other visual record: while this physical action is not excluded from composition, in fact is an integral part of, it is not a focus for investigation. De Beaugrande

(1983a) advocates a 'science of composition': his paradigm envisages a research programme for that discipline which would include: a) a model of the operations and controls involved in writing; b) an account of how writing conditions differ systematically from speaking conditions; c) an explication of strategies of decision and selection; d) a means for decomposing the entire writing process into manageable small subtasks; e) a prediction of the most preponderant difficulties in writing, that is, of the normal weak points in the production system; and f) a set of criteria for evaluating and revising written texts (p.232). It is with the last element in de Beaugrande's programme that we shall be primarily concerned, but clearly work in this area must take place in the context of the discipline as a whole.

1.2. A brief history

1.2.1. Writing, composing and the Greeks

Human communities have developed systems for representing meanings in a more permanent form than sound from the early days of their intellectual development. The mechanical process of learning the written symbols, their significations, and the conventions of their use is an essential step towards writing, but we do not normally refer to this letter-and word-formation process when we talk globally of 'writing'. Rather, we refer to the whole composing process, and its product: the process of selecting from and collocating elements of the written code at a number of levels to form a meaningful whole: a discourse. Our understanding of the art and practice of composing is founded in the language studies of the ancient Greeks, who developed a theory of rhetoric which had two faces: the 'logical', which related to techniques of persuasion and included 'inventio', i.e., invention; and the 'artistic', which related to the aesthetics of style. Aristotle's influence led to the tradition which viewed these two faces of rhetoric as inseparable. Aristotle's work, in particular the first two books of his Rhetoric, were also influential in

that they are considered to have been the first working out of a theory of prose composition: "...we may say that the Rhetoric treats for the first time the art of writing, as opposed to the art of speaking" (Scaglione, 1972: p.19). In both Greece and Rome during the classical period, the study of rhetoric was viewed as a scientific discipline, and engaged the attention of the most prominent creative thinkers.

1.2.2. From composing to composition

Through a long period to the Renaissance, rhetoric remained a dominant discipline in education. Although the art of written composition as a clearly defined part of rhetoric was developed by the Romans, for whom its boundary was the sentence, and its focus was style rather than persuasion, and found its place in the study of rhetoric throughout this period, it was not until the Renaissance that attention was given to the study of prose composition beyond the sentence. This was a consequence of the rationalist influence of Cartesianism and of the Port-Royal grammarians, which caused the stylistic elements of composition to be subordinated to its dialectic function (Scaglione, op cit). Until that time 'rhetoric' had been conceived of as including not only the organization of ideas and their stylistically elegant expression, but also, within the model of classical rhetoric of Aristotle, Cicero and Quintilian, as including an emphasis on constructing persuasive arguments, and particularly on invention. Increasingly, however, concepts of 'invention' were applied to the study of logic and thinking processes. This resulted in the narrowing of attention of 'rhetoricians' towards the composed product rather than the composing process, that is, away from invention and back to style. Increasingly, the concern was with usage (grammar, punctuation, spelling) and with analysis of the product into, for example, discourse types and discourse functions. This was encouraged by the Romantic belief that creative processes are mysterious, impenetrable and therefore unteachable. By the early nineteenth century 'rhetoric' had become 'composition' (Corbett, 1967).

1.2.3. Literacy in the nineteenth century and after

The nineteenth century brought a need for a workforce with some degree of education, and a huge increase in the number of young people benefiting from some form of formal education. The industrial revolution created a need for a class of skilled worker, the first literate working class, who could be trained more efficiently with the help of written materials supplementing the oral tradition of apprenticeship. The spread of the British Empire involved open-sea navigation, which required literate sailors and artillery gunners who could read and calculate angles (Cipolla, 1969). The establishment of the British Empire led colonialists to create a cadre of native clerks and teachers in order to exercise authority in Africa and Asia (McCully, 1965). The working classes were quick to see the connections between literacy and power. From the earliest times, literate members of society have had disproportionate power: priests, oracles, poets and medicine men have all shared as part of their source of power a certain literacy, and in the modern day professors, lawyers, engineers and doctors likewise exercise power in part because of their control over language (Power, 1983).

The view of literacy as an enabling skill fitted well with the view of 'composition' which was current by this period, with its emphasis on usage, grammar and spelling and on stylized methods of paragraph development, focussing on the composed product to the exclusion of invention. A consequence of this view is that the reader/writer has only strictly limited power to manipulate the language to her own ends: the language shapes the would-be member of literate culture rather than the reverse. Bailey (1983) points out that although literacy may create the potential for political transformation, the institutions through which it is transmitted, the schools, "promote the traditional values of a society, foster obedience to authority, and socialize the young to accept roles within the established order" (p. 39).

True literacy must go beyond this. Literacy must be seen not merely as an economic and social survival skill but also as a means for understanding and coming to terms with our lives, as a route to education in the classical meaning, and to full empowerment within one's culture. A modern conception of literacy must see writing and reading as primary tools for the kinds of thinking - abstract, analytic, deductive, inferential - which characterise the most advantaged members of the culture, and to the intellectual and emotional enrichment which almost always accompany such advantage. Such a conception has informed the development of literacy programmes in Britain, the United States, and other industrialised countries since the 1960s, and is increasingly pervading schools and education programmes in Third World countries (Power, 1983). Changing paradigms of 'composition' and 'composing' have played and still play their part in views and values of literacy, and are also affected by changes in views and values of literacy, in a dynamic, continuing process.

1.2. Towards a new paradigm

1.2.1. 'Composition' into the twentieth century

Young (1978) points out that we know less about the development of composition since the nineteenth century than any period before. Kelly (1969) tells us that during this period the formal teaching of 'composition' fell into disfavour and was replaced by translation as the main teaching method for writing in schools: the increasingly sterile view of composition which had emerged toward the end of the Renaissance and particularly in the Romantic period, and the replacement of this view by translation is a further move away from invention and toward style in its most limited interpretation.

It would seem that an impetus for the revival of the use of writing other than for translation came, not from rhetoricians, but from other

disciplines, impelled by a need for a feasible examination system. In the eighteenth century the oral disputation had been the prevailing model of both teaching and testing in the disciplines, but the growing enrolments of the nineteenth century made this increasingly impractical, at the same time that the rapid expansion of knowledge made it difficult (Lunsford, 1986). As the prose composition was increasingly used as the (sole) testing method in school subjects, particularly with the development of public examinations in the mid-1850s (Brooks, 1984), the oral, collaborative model of education broke down and was replaced by an emphasis on the use of writing to demonstrate knowledge. For most of the twentieth century writing has been seen as a means for communicating information, and writing in schools has been primarily taught to enable pupils to show, on written examinations, what they have learned from what they have been taught. Writing was separated from the other communicative arts and lost its purpose as a tool for the pursuit of the individual's academic and social goals, becoming a 'contentless' subject (Lunsford, 1986). The growth of 'objective' testing in the twentieth century is a natural development: if the purpose of writing is simply to demonstrate knowledge, why bother with the writing? Why not get straight to the knowledge?

1.2.2. 'Current-Traditional' paradigm

Young (1978) has described the tradition of composition teaching through the nineteenth and twentieth centuries up to the 1960s as 'current-traditional', and says:

> *The main difficulty in discussing the current traditional paradigm, or even in recognising its existence, is that so much of our theoretical knowledge about it is tacit. (p.31)*

However, he goes on to describe some of its main features: emphasis on product rather than on process; analysis of discourse into discrete, component, parts; classification of discourse into 'types' such as

narrative, exposition and argument; strong concern with usage and style; preoccupation with the formal essay and the research paper.

This paradigm came under increasing attack in the 1960s. At this time, many social forces were on the move: a concern for declining standards of literacy in industrialised countries focussed attention on the teaching of reading and writing in the schools and called accepted models into question. A redefinition of literacy beyond functional literacy to cultural empowerment went hand in hand with the reorganisation of secondary education in Britain. The "60s generation" believed in cultural participation and the ability to affect social change for all members of a social group; writing was seen as a social force, and as a humanizing force. The 'current-traditional' paradigm for the teaching of writing in schools sat uncomfortably with the larger pattern of humanism and concern with the individual, and was criticized for failing to provide effective teaching of invention techniques or of those techniques of analysis and synthesis necessary for the development of thinking, criticisms which culminated in what came to be known as the 'Dartmouth Conference' in 1966. A shift of emphasis in writing research and in writing classrooms to the composing process is one consequence of social change that led to other such shifts, for example, to describing rather than prescribing in linguistics, to learner-centered approaches in education, and to process studies in research in psychology and cognition.

1.2.3. A changing paradigm

Young (1978), drawing on Kuhn (1970), suggests that what has been occuring is a "paradigm shift", and explains this by summarising Kuhn:

> *A paradigm acquires wide support by demonstrating its superior ability to solve problems generally acknowledged by those in the discipline to be acute and fundamental; once it is established, research is directed primarily towards its articulation and application. New problems*

> *arise, however, which those committed to the paradigm cannot solve adequately, and a crisis develops, accompanied by a sense of uncertainty and insecurity in the profession. The response to the crisis is typically the development of new theories which are able to provide more adequate solutions. A new paradigm emerges from the inquiries and controversies of the crisis state and with it another period of relative stability. (p.35)*

Hairston (1982) suggests that external circumstances hastened the sense of crisis: she identified open admissions policies; the national (U.S.A.) decline in conventional verbal skills; the increasing proportion of high school graduates entering tertiary education; and the entry to tertiary education of increasing numbers of armed service veterans and other older, more demanding, students. Similar external circumstances in Britain may have been the raising of the school-leaving age and the introduction of the C.S.E. examination. Hairston describes several *ad hoc* remedies which were tried in the period of change, among them writing labs, individualised instruction, expressive writing and sentence combining. She believes that each of these contributed insights but none was important enough to indicate that a true paradigm shift was needed, and she cites the work of Shaughnessy (1977) in the U.S.A. and Britton et al (1975) in Britain, both of whom began their studies in the late 1960s. Another important influence was the work of Murray (1968), which challenged many assumptions about the traditional approach to the teaching of writing and emphasised writing as a process of self-discovery.

In 1963, Braddock, Lloyd-Jones & Schoer said that "today's research in composition ... may be compared to chemical research as it emerged from the period of alchemy" (p.2). In the intervening twenty years, there have been signs that we have indeed emerged from the 'alchemy' of composition and, while we may not quite have a periodic table of the elements of composition, we are certainly in a much better position to be able to understand and describe, from empirical studies, the nature of writing

development, the components of the composing process, and the features of 'good' writing.

1.2.4. 'Process-Invention' paradigm

Young (op cit) saw two important changes which characterised the new paradigm: a shift in attention from composed product to composing process, and a revival of interest in and teaching of invention.

Studies such as those referred to above, by Shaughnessy and Britton, describe the learner-writer at developmental stages and question why the writer is as she is. Other studies, such as Emig (1971), Perl (1979), Graves (1983), and the work of Wilkinson and his team on the 'Crediton Project' (1978; 1978; 1980), have helped us to understand how the young writer develops towards maturity in composing. Weaver (1973) explored similar questions with English teacher candidates. The development of self-report techniques and protocol analysis by Flower & Hayes (1977,1979, 1980a, 1980b) and a coding system for protocols by Perl (1981) has provided a research methodology and permitted detailed and structured observation of the writing process. Studies by Hunt (1965; 1970; 1977), Witte (1980; 1983) and others on T-unit length and other syntactic maturity measures have quantified certain features of writing development. Research in educational psychology such as Peel (1971), Marton and Saljo (1976), Biggs and Collis (1982), and a renewed interest in the application of Piaget's work have also made contributions to the methodology and hypotheses for investigating the composing process.

Models of the composing process, on the lines of Cicero's invention, arrangement, style, memory and delivery model, have been developed. For example, Bruce et al (1978) produced a stage model consisting of: discovering ideas, manipulating ideas, producing text, and editing text. Britton et al (1975) propose a three stage model: preparation, incubation, and articulation. Murray's (1978) three stages of prevision,

vision and revision are very similar. De Beaugrande (1983) proposes a 'multi-dimensional parallel-stage' model of the writing process: he believes that all 'stages' operate for each stretch of text, with dominance shifting among them over time. He says:

> *One conclusion of the multi-dimensional parallel-stage model is that text production has no clearly built-in conclusion, no point at which cognitive processes are definitely accomplished ... (the model) implies stages, thresholds, shifting back and restarting/revising. (p.237)*

Flower and Hayes (1980) also refute a linear view of the stages of the composing process. They propose that writing consists of three major processes: planning, translating and reviewing. Planning consists of the sub-processes generating, organizing, and goal-setting; reviewing consists of the sub-processes reading and editing. Flower and Hayes' model is dynamic, and they believe that the writer moves through the processes/sub-processes constantly and in rule-governed ways. For example, editing has priority over all other processes and can interrupt any of them. Using think-aloud protocols for the observation of the composing process they have established five principles which are empirically supported:

1) writing is goal-directed;

2) writing is hierarchically organised (i.e., goals are set; sub-goals are set in order to achieve goals; etc);

3) writing processes can interrupt other writing processes over which they have priority (see above);

4) writing is a recursive process (i.e., it contains each of its parts within its smaller parts; for example, an 'edit' interrupt sets in motion the whole set of processes within the 'edit' sub-process);

5) goals can be modified during the process (Flower and Hayes have not determined where goal-modification fits into the model: they suggest as part of the 'goal-setting' sub-process).

Black et al (1983), describing what good writers know, talk of goals, plans, scripts and themes. Goals are those of the writer, and of characters in a story or a play (essentially the sense is the same as in Flower and Hayes, above). The making of plans for the writing, and the carrying through of these into discourse realisation, is seen as a method for accomplishing a goal: while plans are abstract, scripts are a routine method of accomplishing a goal. For example, we share conventions of behaviour for entering a restaurant, ordering, and paying the bill: the use of this convention in a written discourse is a script. Themes are the elements of background knowledge which make it possible for the reader to predict the writer's goal.

Awareness of the importance of the reader is one of the most important characteristics of the good writer. Flower (1979) describes what she calls 'writer-based prose' as a halfway stage in the composing process, in which search and selection procedures are mainly complete and appear as a "rich compilation of thoughts" which, as long as the audience is the writer, constitute well thought-out communication. However, she describes writer-based prose as unsatisfactory for any other reader due to such characteristics as missing information, lack of organization, and the omission of syntactic elements, particularly psychological subject. The good writer goes on to transform this writer-based prose into reader-based prose, i.e., into an autonomous text. Nystrand (1983b) has pointed out the importance of this central awareness of the reader for good writing:

> *When written communication fails, readers find the text misleading, turgid or ambiguous. Aware of text rather than meaning, these readers are in effect excluded. By contrast, when written communication occurs, readers find the text legible, readable, and lucid - in short, "transparent". Unaware of text as text, they are "absorbed" into the world of its meaning...This transformation points to a confluence of reader-writer consciousness - its effect underscoring their participation in a shared space - textual space. (p.72)*

Young et al (1970) describe how difficult it is to achieve this participation in shared space:

> *...there can be no interaction between writer and reader, and no change in their thinking, unless they hold certain things in common, such as shared experiences, shared knowledge, shared beliefs, values and attitudes, shared language. (p.172)*

They relate this to Grice's (1975) maxims, springing from the cooperative principle. Miller and Kintsch (1980) are in agreement:

> *Readability is not a property of a text... (it) is an interactive relationship between the properties of a text and the reader who is processing it. (p.348)*

This is a view central to the 'reader-writer contract' proposed by Tierney & LaZansky (1980), in which both writer and reader understand and accept that they are co-signatories to an agreement with firm and mutually binding conditions. The implication of this centrality of the reader to the effectiveness of any written discourse is that any research into composition must include a consideration of the reader and his responses, as Nystrand (1983c) points out. Flower and Hayes (1980) describe three constraints on the (adult) writer of expository prose: firstly, they need integrated knowledge (i.e., a conceptualized and precisely organised knowledge network); secondly, they are constrained by the linguistic conventions of written text (i.e., explicitness, cohesion and cohesion, lexical fluency, etc.); thirdly, there is the rhetorical constraint (i.e., the writer must conform to the structures imposed by purpose, audience specifications, and writer's roles). They stress that composing is a speech act: because writing is not supported by an actual context, by the existence in the same space-time dimension of those concepts or objects to which it refers, it must be independent, providing all its own referents.

The view of composing as a speech act returns us to the second of Young's characteristics of the new paradigm: invention. A key feature of the new paradigm as it is manifested in writing classrooms is an attention to prewriting as a means of discovering meaning, i.e., what it is the writer has to say. Invention techniques play a major part here. Young describes four main approaches to invention: classical invention (developed by Aristotle); Burke's dramatistic method (heuristic probes); Rohman's prewriting method (journal writing, pseudo-religious meditation techniques and analogy); and Pike's tagmemic invention (problem investigation and solution). Some other invention techniques are cubing and looping (Cowan and Cowan, 1980).

1.2.5. The present position

Currently the field of composition research and teaching is emerging from a period of wholesale acceptance of this new paradigm and into a more mature period, in which a focus on writing as a process no longer excludes all attention to the composing product. Hillocks (1986) conducted a meta-analysis of all published studies from 1963 (the year of the publication of Braddock et al's survey) to 1982, finding a good deal of support for the process paradigm, but also some unsupported assumptions and some contra-indications. Many studies showed that the writer's process is influenced in important ways by external factors, even from the earliest ages. Hillocks categorises three 'modes of instruction' as a result of his meta-analysis. The 'presentational' mode represents the established tradition of composition teaching, and of much instruction in other areas also. In this mode the teacher dominates all activity and learners are passive recipients of knowledge. He found that was the least effective mode of instruction. The 'natural process' mode as it occurs in writing instruction can be dated from the Dartmouth Conference and the work of Emig (1971), and Hillocks describes it as:

> *"a reaction against the dominant presentational mode with its often arbitrary assignments given with no preparation; with its structures to be learned from rigid models, such as the "five-paragraph theme"; and with its emphasis on the "correctness" of products. (p.247-8)*

The third mode, the 'environmental' mode, appears to be relatively recent, although Hillocks sees its intellectual roots in Herbart and Dewey. The environmental mode moves beyond process without abandoning it; for example, proponents recognise the need for prewriting, but they focus on prewriting activites which help develop skills which can be used in composing products. Environmental instruction may use models and teach form, but it emphasises activities which help writers understand the thinking which lies behind the forms and encourages them to look critically at models. By incorporating aspects of both earlier modes (which bear striking resemblances to the paradigms we have been discussing in this chapter) it moves beyond both to a more powerful paradigm.

1.2.6. From L1 to L2 composing

The discussion of what we know and believe about the composing process in the L1 has implications for a better understanding of composing in the L2. The principles which can be derived from the research discussed above provide a theoretical underpinning from which L2 composing research can draw hypotheses and procedures. In particular, L2 composing research has to orient itself towards the investigation of the relevance of the various paradigms discussed above for L2 contexts. What Cooper and Odell (1978) said in relation to L1 writing research has until recently been true of L2 writing research:

> *For too long a time, many researchers assumed that the most significant kind of question was: What materials and procedures will improve students' work in written composition? Underlying this question was a further assumption that we did, in fact, have an adequate understanding of the term 'composition', that our primary*

> *job was determining the effectiveness of specific instructional materials and procedures, rather than finding out exactly what information and skills teachers and researchers ought to be concerned with. (p.xi)*

The traditional approach to teaching second language writing, like the current-traditional paradigm in L1 writing, has depended on teaching the 'grammar of writing' and the rhetorical-structural conventions of written text: this has been particularly true for the teaching of expository writing, where this has been taught at all. The process-invention paradigm in L1 writing has had a tremendous impact on the teaching of second language writing, in showing that there is a coherent body of knowledge about how writers write and in making suggestions about how writing can be taught taking into account the new insights into the cognitive and psychological processes involved in composing text. However, these developments are less well known outside the English L1 countries and will take many years to have significant influence. There are also indications that a third approach, which like Hillocks' environmental mode combines the best of both earlier approaches, is gaining acceptance among L2 writing researchers.

1.3. Composing in a second language

1.3.1. L2 studies of composing: focus on products

Fein (1980) compared native English and ESL student writers in equivalent undergraduate writing courses and found that although the ESL writers recived much lower ratings on impression marking and on an error count, they compared more closely to the native English writers on content, organisation and style. He suggests two reasons for the differences he observed: the language acquisition process may be at work, and fluency, rhetoric and grammatical complexity may be acquired before grammatical accuracy; or errors may have been fossilized. However, the raters he used were the English class teachers, who did not know that ESL essays were

included in the papers they were marking, and they used their own criteria. As is shown in Chapter 2, section 3, the orientation of the rater, and the criteria used, play an important part in the resulting assessments. It may have been that Fein's raters were heavily error-focussed: unfortunately Fein does not provide sufficient information on these aspects. He concludes that, because ESL writers have some grasp of the 'five paragraph' fundamental structure of an expository essay, teaching should not concentrate in this area but in the areas of weakness revealed in his study, i.e., grammatical accuracy.

Edelsky (1982) studied the writing in English and Spanish of bilingual primary children. She collected pieces of writing from three class levels at different times during a year and studied them in an attempt to understand the relation between the children's writing in L1 and L2. She found that the children applied high level strategies from Spanish (their stronger language) to their English writing but that these were affected by what the children knew about English. Edelsky concludes that "what a young writer knows about writing in the first language forms the basis of new hypotheses rather than interferes with writing in another language" (p.227), and suggests that application of L1 knowledge to L2 writing can appear in both similarities and differences in texts in the two languages by the same child.

Jacobs (1982) investigated the writing of five ESL writers and six native speakers of English, all of whom were, however, from ethnic minorities. All eleven writers were grammatically competent; they were all in the same pre-medical program. Each student wrote on a topic from their lecture course each week, and the writing was responded to by Jacobs, who sat in on the course with them. All the writing tasks were of a formal, expository nature, and the writing was done in an examination-type situation. Jacobs assumed that all the students had mastered the content, and wanted to force them to use organizing principles to present their writing; although the students could refer to their notes, she

ensured that the questions could not be answered using the organizational structure of the lecture(s) in which the notes had been taken. She concludes that academic writing demands high predication loads (that is, it is of high rhetorical and relational complexity), and that because high predication load is difficult to handle it leads to less coherent text. However, these problems are not noticeable until the writer is in control of the content; if she is not, she will have few propositions to build into predications in the first place. When students had integrated the content and were seeking to organise it appropriately, their phrasing changed for the worse, showing more grammatical faults. Those students who had not yet grasped the need to integrate the new content into their own thinking and to represent it in a modified/selected form to fit the task set did not show these changes for the worse in phrasing. Jacobs believes that this shifting level of grammatical control is symptomatic of the developing writer. Further, she believes that the same writer may on some occasions be an integrator of information and on others simply represent the original information almost regardless of the task: this is another aspect of the development of the writer as a thinker, since in Jacobs' view writing is a formal discipline which through its exercise brings learning. Jacobs also found that the ESL and native speaker writers showed the same problems and strategies.

These studies all approached L2 composing through products: more recently, other studies have focussed on processes.

1.3.2. L2 studies of composing: focus on processes

Lay (1982), for example, had students compose aloud while she observed them, and concluded that most of the processes observed in studies of L1 composing are also present in L2 learners. Zamel (1982) interviewed eight skilled ESL writers, and examined their writing at various stages: she found that they used similar strategies to those used by L1 student

writers. In a further study (1983) Zamel looked instead at the composing behaviour of six advanced ESL writers, and concluded that:

> *...instructional approaches that view writing as the sequential completion of separate tasks, beginning with a thesis sentence and outlines and requiring topic sentences before one has even begun to explore ideas, may be as inappropriate for ESL students as they are for native speakers of English. (p.181)*

Heuring (1984) observed the revision strategies of five ESL writers at different proficiency levels, using a video-taping method that allowed him to collect both process and product data. Heuring makes three general observations from his data: 1) skilled writers gave revising a complementary and productive role in the writing process while unskilled writers were not able to strike a balance between revising, planning and transcribing; 2) skilled writers revised at the level of meaning as well as of surface features, while unskilled writers were preoccupied with revising local, surface-level features; 3) the most skilled writer used reading as a revising strategy to a considerable extent, while the least skilled writer did not use any in-process reading strategies, re-reading only between drafts. Heuring's observations led him to suggest that more emphasis in teaching should be placed on idea generation and development than on grammatical accuracy.

Raimes (1985) asked relatively unskilled ESL writers to compose using the think-aloud protocol technique and Perl's (1981) coding method, during their normal ESL writing course. Raimes found that her students not only attended to the task but were "riveted on it" (p.246); unlike Perl's (1979) writers, they were not preoccupied with errors or editing. Also unlike Perl, Raimes found few common patterns among her eight writers, but she did find that all eight were able to generate language and ideas in much the same way as more proficient students. These writers did not view writing from the reader's point of view, unlike Zamel's (1983) skilled ESL writers who "understood the importance of taking into account a reader's

expectations" (p.178). Raimes concludes that ESL students in writing classes cannot be treated in the same way as L1 student writers: while attention to process is necessary, it is not sufficient:

> *What the less proficient writers need is more of everything: more time; more opportunity to talk, listen, read, and write in order to marshal the vocabulary they need to make their own background knowledge accessible to them in their L2; more instruction and practice in generating, organizing and revising ideas; more attention to the rhetorical options available to them; and more emphasis on editing for linguistic form and style.*

The view Raimes puts forward in this article can be seen as a call for a third approach, where the emphasis is exclusively on neither processes nor products, but on both in the kind of mutuality described by Hillocks (op cit) as characterising the environmental mode.

1.3.3. Research on L2 composing: competing paradigms

The literature on L2 composing from the end of the 1970s has been dominated by proponents of the process-centred approach, but this approach has not yet shown that teaching writing within this paradigm to L2 learners actually leads to better writers: we do not yet have research evidence that emphasis on process leads to a better product in L2 classes, as Horowitz (1986) points out. Yet wholesale criticism of the current-traditional approach is common. Taylor (1981), for example, criticizes the practice of the teaching of outlining and the use of models, and claims that revision is "largely unexplored in most writing programs" (p.7). Taylor says:

> *Rather than offering students assignments which require that they grind out essays on teacher-assigned topics on the spot, or imitate a model, or follow a controlled exercise, it is more effective to teach students to build up from their own written ideas. The notion of revision could hardly be made more explicit than by having students sit down with their own random, isolated sentences and*

> *phrases from their journals, debates, and brainstorm sessions and begin to pull them together. (p.11)*

However,Taylor does not provide data from L2 studies to show that these classroom practices result in measurable improvement in ESL students' writing. Zamel (1982) has a similar position, and concludes:

> *If, however, students can learn that writing is a process through which they can explore and discover their thoughts and ideas, then product is likely to improve as well. (p.207)*

It is clear that the new paradigm has demonstrated its superiority to the old in humanistic terms and in the way it informs and is in turn informed by interlanguage studies. Few teachers who have made the conversion in their teaching from a product-centred to a process-centred approach have failed to discover for themselves its superiority in terms of student involvement and interaction and therefore of motivation. Similarly, the techniques for the generation of ideas, the treatment of error and the approaches to feedback in a process-invention approach are appealing to both teachers and learners. But some proponents of a process orientation in the teaching of L2 writing (notably Zamel and Taylor, above) have rejected any role for attention to products, and have discounted the relevance of product requirements in educational systems. What is needed, however, is research rather than polemic and hypotheses: without the results of such research are available, the process approach is as vulnerable to assault as the product approach has been.

A lead has been given, although of a limited kind, by Spack (1984), who presented a case study of a Paraguayan first year undergraduate who was taught, and used, invention techniques in his English composing. Invention, or discovering what one knows, how one feels, and what one wants to say about a topic, has a central place in the new paradigm and is a specific classroom technique which can be applied and observed.

Spack believes that through using invention techniques, the student she observed:

> *"discovered an organic relationship between the content of his subject and its form...(he) made an effort to synthesize and further clarify his ideas, to put them down in an organized fashion, to include examples relevant to his audience, and to construct well-developed sentences. A concern for organization and correctness gained priority in his writing after ideas came to life in invention". (p.662)*

Selinker and Kumaradevelu (1986) describe a 'safe-rules' approach to L2 composing, in which writers are taught strategies for composing which are safe,and alternatives which are less safe, to be used as the learner increases in confidence. Swales (1984) presents a heuristic for the composing of introductions to research articles, and in a forthcoming paper he shows how the heuristic was taught to and applied by a group of graduate engineers with some success. Swales' heuristic combines features of both the process paradigm and of the product paradigm.

Spack (op cit) concluded that the teaching of writing in a second language cannot depend for its techniques on the findings of L1 studies, and that research into teaching practices which will uniquely suit L2 students is sorely needed. Until the results of such research are known, there will be competing paradigms and little more than individual intuitions to guide choice between them. There have already been indications, for example by Horowitz (op cit), that some L2 teachers are unhappy with the strong view of the 'process-invention' paradigm and feel a strong need for evidence of the benefits for learners in instrumental terms. If it is true, as Horowitz claims, that the process approach only allows for certain ways of seeing, thinking and writing, then the process paradigm will have fallen into the same traps as the product paradigm it has superseded. Although there is no L2 equivalent of Hillocks' L1 meta-analysis, Hamp-Lyons (1986a) suggests, and Horowitz (1986b) concurs, that it is essential to reconcile differences among proponents of one paradigm

or the other and to move forward to a descriptive model, based on research into both writers' composing processes and their written products, which will be more powerful

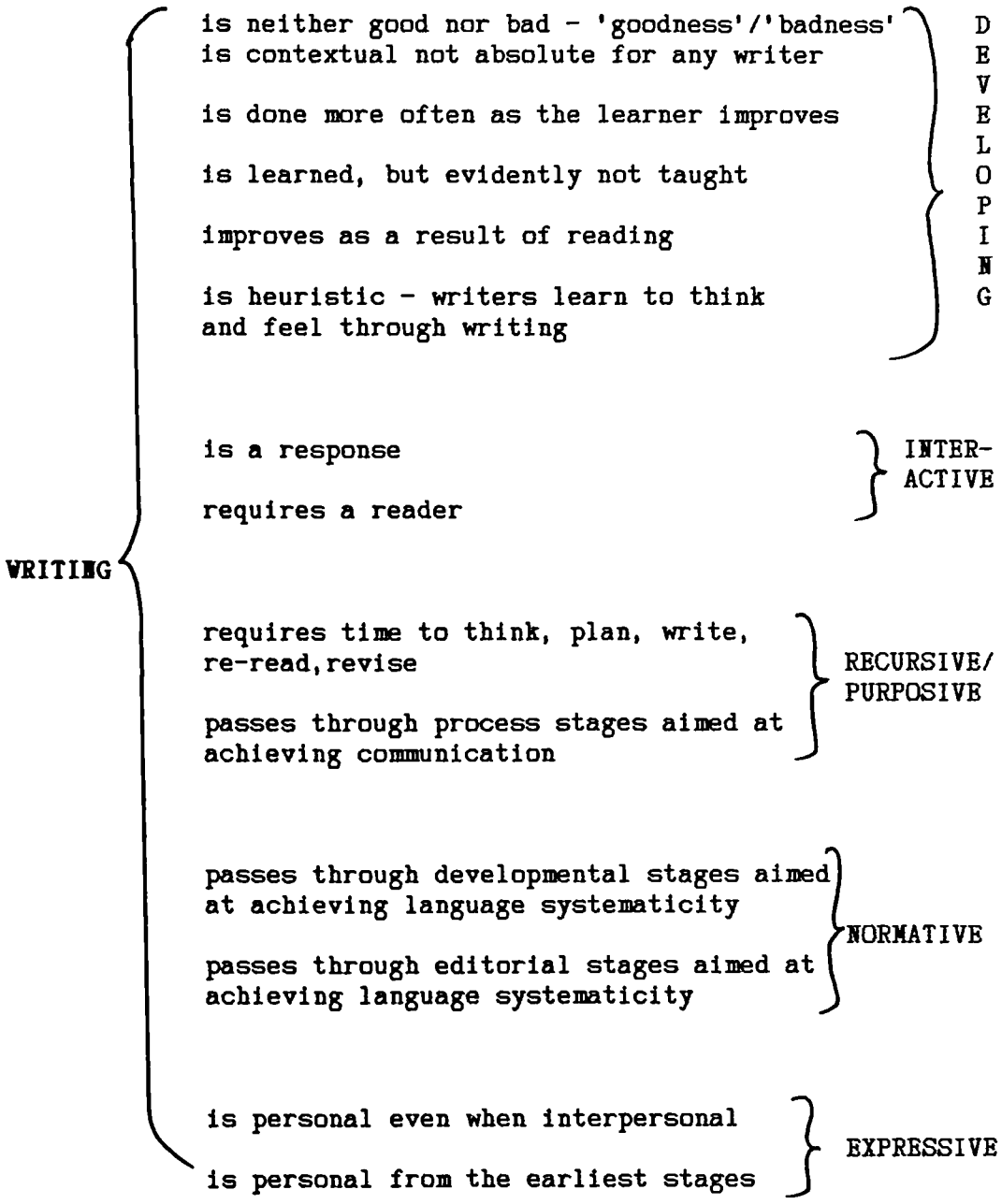
1.4. Research on L2 composing: implications for testing writing in academic settings

It can be seen that the limited amount of research which has been done in L2 composing, and the fact that most of this has focussed on writing processes without regard to absolute judgements of the resulting products, leads to problems for the evaluation of second language writing in academic settings, whether specific or general academic. We are not yet in a position to claim that we have a construct of either the processes used by unskilled, semi-skilled and skilled L2 writers when they compose in a highly structured context on an assigned expository topic or of the standards by which the products of these composing activities should be judged. As Nystrand (1983b) points out:

> *Any valid and useful assessment of writing must cope with enormous problems entailed by the absence of an abiding, lawful account of how writing works. For the test maker, it is an issue of construct validity in the absence of a construct. (p.18)*

The diagram below (Figure 1.1.1) is an attempt to distil the findings of the survey and discussion in this chapter into a brief characterisation of the composing process which can be applied to the investigations of the testing of writing in academic contexts in Chapters 4, 5 and 6.

Figure 1.1.1


WRITING
is neither good nor bad - 'goodness'/'badness'
is contextual not absolute for any writer
is done more often as the learner improves
is learned, but evidently not taught
improves as a result of reading
is heuristic - writers learn to think and feel through writing
DEVELOPING
is a response
requires a reader
INTER-ACTIVE
requires time to think, plan, write, re-read, revise
passes through process stages aimed at achieving communication
RECURSIVE/ PURPOSIVE
passes through developmental stages aimed at achieving language systematicity
passes through editorial stages aimed at achieving language systematicity
NORMATIVE
is personal even when interpersonal
is personal from the earliest stages
EXPRESSIVE

2. CONCEPTS IN LANGUAGE TESTING

2.1. Language testing defined

We find in the literature of the field a number of terms which are generally used without definition, but which are intended to convey meaning. 'Testing' is itself a term of this type. Lado in his classic work (Language Testing: 1961) defines language but not testing. Valette (1977) similarly does not provide a definition for the global term 'testing'. Davies (1968) comes closer, with his discussion of the relation between theories of language and learning, and testing, but does not provide us with a precise definition. Heaton (1982) and the contributors to his volume similarly do not provide us with a definition, though Heaton's statement that "the whole psychometric basis of language testing has been seriously questioned" informs us at least that there has been a psychometric basis for language testing. There seems, then, to be a tradition of using the term 'language testing' undefined, and Stevenson (1981) provides us with clues to explain the non-definition of language testing with his discussion of the restricted and expanded views of language testing, and of the conflicts generated among proponents of the different views, as well as between language testers of any kind and the lay person who is affected by work in language testing.

There is, perhaps, a folk wisdom in leaving language testing undefined, for in defining one's ground precisely one takes on the obligation for defending it territorially, and loses the possibility of shifting it. The current period in language testing does not seem to be the time or place to set up camp and defend a position against all comers, as Oller's experience has shown (1983). In the 70's Oller had taken up a well-defined position and defended it consistently: in his more recent work (op cit) he has retreated from that ground, albeit graciously, and has joined the majority of language testers in the open forum. As Stevenson (op cit) points out:

> *...views of language testing are not available as pre-packaged and competing credos in the testing literature. Rather, they represent fluid concerns and sets of emphases." (p. 17)*

Language testing, then, means different things to different people, and until we can all agree on a common definition of the term, any unilateral attempt at definition will only lead to confusion and arouse hostility. Common practice is followed here in using the term without defining it.

2.2. Expectations language tests must fulfil

The expectations which are placed on language tests are dependent to some extent on the purpose for which the testing is being done and the strengths of the claims which are made for the test instrument used. Clearly, a standardized placement instrument used nationally or internationally must meet more stringent criteria than an *ad-hoc* classroom test used only for internal diagnosis/remediation. The expectations of language tests are the same as for all kinds of tests: reliability, validity, practicality, and positive backwash.

2.2.1 Reliability

Carmines and Zeller (1979) define reliability as "the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials." (p. 11). No measuring procedure, in any discipline, yields perfectly reliable results, and the finer the measure which is applied the more measurement discrepancies will emerge. We cannot expect duplication of results, but we can expect consistent results. However, any test will contain random error (i.e., chance factors which interfere with precise measurement), and random error is inversely related to the degree of reliability of the test. Test constructors aim to reduce random error to a minimum. The amount of random error which can be tolerated in a test depends on the stringency of the expectations for that test.

Testing experts do not agree on the tolerance limits for random error. Ingram (1977) expects properly constructed large-scale tests to have a reliability coefficient of at least .95. Harris (1969) believes that any standard test designed to separate one examinee from another should have a reliability coefficient of at least .90. Carmines and Zeller (op cit) set a lower limit of .85 in similar situations. Thorndike and Hagen (1969) remind us that there is no fixed minimum reliability required of a measuring instrument, but that, other things being equal, we should select the instrument with the highest degree of reliability. What is always possible is the specification of the level of accuracy attributable to a description of any individual on the basis of a test score, and this specification should always be provided. Guilford and Fruchter (1978) similarly remind us that reliability must be interpreted in a relativistic manner.

There are two ways of establishing the reliability of a test: through measurement of stability, or through measurement of equivalence. Determining stability reliability involves administering the same test to the same students with a time gap between. The subjects are expected to retain the same rank order from test to retest. The problem with using test-retest as a reliability measure is that human subjects cannot be depended on to perform consistently. In addition, it is impossible to control what the subjects do between the two administrations of the test: some subjects may receive intensive instruction which will affect their performance on the retest, for example. Another way of establishing stability reliability is to mark the same test more than once. There are two approaches to determining mark-remark reliability: firstly, the same examiner marks the same test papers on two occasions with a time gap between; secondly, two or more examiners mark the same test papers. Whichever method is used, what is being investigated is the extent to which the test instrument is consistent in its measurements. There are several ways to determine equivalence reliability. A test-retest method in which parallel forms are used, rather than the same version,

establishes the extent to which the two sets of scores are measuring the same thing. However, aiming at equivalence reliability by parallel forms is a circular operation in that it seeks to determine the reliability of the two tests by comparing them with each other, the reliability of neither necessarily having been already established externally. It is only useful when Form B of a test is being tested for equivalence reliability against Form A, reliability of which has already been independently established.

The Spearman-Brown split-half technique resembles the parallel forms method, except that a single form of the test is divided (split) into two halves, usually odd and even, and the scores are treated as parallel forms. The two disadvantages of this technique are that there are fewer items than in the parallel forms methods, and that, as Richardson and Kuder noted (1939), the value of the reliability coefficient thus obtained is not unique. The assumption underlying the Spearman-Brown technique is that both halves of the test will have identical standard deviations, but in fact the specific split chosen will determine the level of equivalence reliability obtained.

Kuder and Richardson (1937) developed an alternative technique for establishing equivalence reliability, in which the coefficient of equivalence is defined as the relationship between one form of a test and another, hypothetical form. Equivalence is defined precisely in terms of the items or elements in the test, and departures from exact equivalence are directly measurable. There are four formulae, all of which can only be used with dichotomously-scored items. The simplest formula, known as KR 21, can be used where there is good item homogeneity and where a conservative estimate of variance is acceptable. An alternative formula, KR 20, is used when there is doubt that item difficulty is consistent, and when a more precise measure of variance is required.

Heaton (1975), discusses five factors which may affect the reliability of a test:

> 1. *size of the item sample selected for testing (the greater the number of items, the higher the chance of reliability);*
>
> 2. *test administration (conditions should be the same for all testees on all occasions);*
>
> 3. *test instructions (do all candidates understand what is expected of them?);*
>
> 4. *personal factors (e.g. motivation, illness);*
>
> 5. *scoring (objectivess tests build in marker reliability; subjective tests do not necessarily do so). (p. 155 - 161)*

It can be seen that techniques for determing test reliability centre on (1) and (5) above; test administration and personal factors are not susceptible to either objective study or control by the test constructor or validator. Although care is normally taken to provide clear test instructions there has been little investigation of the effect of test instructions on a test's reliability.

Reliability refers to the consistency with which a test measures whatever it measures, and it tells us nothing about what the test actually measures. Harris (1969) points out that no matter how high the test's reliability coefficient is, "it is by no means a guarantee that the test measures what the test user wants to measure" (p. 18). Yet as Davies (1977) says "...unreliable results can have no meaning apart from their own randomness. ...it is essential to establish reliability first. Otherwise there is no point in considering validity" (p. 57). There is a problem inherent in the relationship between reliability and validity to which we shall return after discussing validity in detail.

2.2.2 Validity

In the field of language testing, validity has traditionally been defined according to Lado (1961): " ...it answers the question 'Does the test measure what it is intended to measure?" (p. 30) Davies (1977) calls this the "common sense" approach, since it relates the test to a pre-determined purpose. The problem, he says, lies in assuming that the exact purpose of a test can be known: for this a criterion is needed. The criterion is a representation, or statement, or alternate measure, of what the test purports to measure. Validity questions, then, lead inevitably to questions as to what it is which is being measured, or which should be measured (p. 58).

Cronbach (1971) makes it clear that validity is not a simple, unidimensional concept: "Validation of an instrument calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate. The investigations supplement each other." (p. 445). In identifying the types of validity which a language test should possess, it is most common to follow Cronbach (op cit, p.106), who lists four: predictive, concurrent, content, and construct. The American Psychological Association, in association with the American Educational Research Association, combined concurrent and predictive validity and established three aspects of validity which should be researched before any test is widely distributed: criterion (= concurrent + predictive), content and construct (APA,1966). Criterion validity is also referred to by Harris (1969) and others as empirical validity.

Whichever classification of the types of validity is used, it remains the case that validity concerns the crucial relationship between a concept and an indicator of that concept: in assessing the validity of any measure there are always theoretical claims being made about this relationship (Carmines and Zeller, 1979). An additional type of validity

which is considered to be important by most language testing researchers is face validity, which is generally accepted as one of the 'five validities'.

2.2.2.1 Face validity

Face validity is rarely referred to in the classical literature on psychological testing from which the bases of our thinking on language testing are derived: it does not appear in the APA's standards (op cit), for example, nor in Cronbach's classification. Davies (1977) explains that this is because it is not a theoretical concept. Face validity refers to the degree of acceptability of the test to the lay eye. Morrow (1977) explains this by saying "...it must seem plausible to the person taking the test that the tasks he is asked to undertake are relevant to the objectives of the test" (p. 16). Hughes (1981) points out:

> *"A test's lack of face validity will have a detrimental effect on predictive or concurrent validity; at least some candidates will fail to take the test seriously, and so their performance on the test will not provide an accurate picture of their ability." (p. 208)*

However, Stevenson (1985) warns that all testing contexts are 'inauthentic' in that behaviour is required and observed for scoring purposes and not for real-life purposes and thus possess limited face validity. Palmer and Bachman (1980) refer to face validity as "the least important type" of validity (p. 1). However, the importance of face validity should not be underestimated: it must be remembered that the lay persons who will judge a test on its face validity include not only the testees, but also administrators deciding among a range of available test instruments for specific purposes of their own, often without consulting language testing specialists; and score consumers who may use scores on a test instrument for making decisions about individuals based entirely on their perception of the purpose and meaning of the test and individual scores attained on it, i.e., entirely on face validity. Clearly these are

both serious considerations, despite (or because of) their lack of an adequate theoretical base. However, Stevenson (op cit) warns language testers against using face validity as a criterion themselves: discussing a study of different methods of assessing ESL writing ability, he says:

> *So-called direct approaches, that is, an essay and a reader...are favoured. However, it is admitted that many criterion-related studies do seem to show that so-called 'indirect, objective' tests are relatively valid. And also, they do have several psychometric advantages over direct essay grading as it is usually carried out in the real world. But, nonetheless, the conclusion is given that they cannot be valid, 'they lack face validity...and construct validity' (Stevenson's emphasis). This is because, of course, we know they can't really be measuring what they appear to be. (p.43)*

2.2.2.2. Criterion validity

Cronbach (1961) identifies two types of criterion validity, concurrent and predictive; and distinguishes between them according to whether test scores are compared with a direct measure of the criterion performance collected at virtually the same time (concurrent), or whether test scores are used to predict a future criterion and compared with that criterion at a future date (predictive). Davies (1983) does not wish to distinguish rigidly between the two because "... for me the issue is largely practical in terms of when the criterion is available for observation and measurement"(p.142).

Carmines and Zeller (op cit) consider the logic and procedures to be the same for concurrent and predictive validity, the only difference being whether the criterion variable exists in the present or the future. However, this seems to ignore Cronbach's view that the purposes of determining concurrent and predictive validity are essentially different: he gives the principle use of concurrent validity as the substitution of a more for a less convenient procedure, and the principle uses of

predictive validity as selection and classification of testees (op cit). Anastasi (1976) makes a similar distinction.

2.2.2.2.1 Concurrent validity

Cronbach (1961) states that concurrent validity is investigated when a new test is proposed as a substitute for some other information. He suggests that this occurs when a test of a certain ability involves an inconvenient procedure, and the new procedure is expected to give an equally acceptable estimate of the same ability. In such cases the previous procedure is used as the criterion for estimating the acceptability of the new test. Morrow (1977) states that concurrent validity is established when scores on the new test correlate highly with scores on existing tests whose validity has already been established (p. 16). This, however, leads us to a problem. How is the validity of the earlier test established? Davies (1977) pinpoints the danger of regressiveness, i.e., "...if test X is established by concurrent validity on test Y which was itself validated against Z, then a certain drift is engendered" (p. 60). For this reason, Davies suggests that concurrent validation should also look outside existing tests for criterion measures. Ingram (1977) found teachers' ratings to be one of the best criterion measures of a test's concurrent validity. In developing the first version of the English Proficiency Test Battery (EPTB), Davies (1965) used teachers' ratings on a variety of scales as criteria.

2.2.2.2.2 Predictive validity

Pilliner (1968) describes predictive validity as "a numerical expression of the correspondence between performance on the examination used as a predictor and some criterion of later success" (p. 87). Morrow (1977) adds that predictive validity means the test will predict successfully the performance of students at a later date (my emphasis). His example is the EPTB. The greatest problem in predictive validity, according to

Thorndike and Hagen (1969) is finding a satisfactory criterion. They point out that any realistic criterion can only be partial: the ultimate criterion will be lifetime success in the sphere for which prediction is made. There are many potential criteria against which to measure the predictive validity of any test instrument: success in a job or training programme; supervisors' ratings: teachers' ratings; colleagues' ratings; scores on later tests; academic outcomes. The characteristics recommended by Thorndike and Hagen (op cit, p.168) to be sought when choosing a criterion are (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability.

Predictive validity coefficients are typically rather low: the highest quoted by Thorndike and Hagen (op cit, p. 170) is .78 for Lorge-Thorndike Verbal Intelligence Test with Iowa Tests of Basic Skills, with other coefficients ranging between .01 (Differential Aptitude Tests - Space Relations with English grades 3½ years later) and .60 (Short Employment Test - Arithmetic with stenographers' job grade). Some of the reasons for low predictive validity coefficients are: time lapse between predictor and criterion; unreliability of the predictor; unreliability of the criterion; lack of relationship between predictor and criterion; individual aptitude for criterion as opposed to predictor. Davies (1965) reported a predictive validity coefficient of .45 for his EPTB with academic outcomes. Burgess and Greis (1970) looked at scores on various language courses and on the Michigan Test, the Lado Test, and TOEFL (Test of English as a Foreign Language) to find the best predictor of Grade Point Average (GPA: success-rate session by session in American university classes) and Cumulative GPA (i.e. success level at the end of degree): the best predictor was the Writing Grade (English 110) at .70 with GPA and .62 with CGPA. Heaton and Pugh (1974) found a predictive validity coefficient of .22 for the University of Leeds English Test with academic outcomes. Chai and Woehlke (1979) discuss the predictive validity obtained in a number of studies comparing American ESL tests used as college/university admissions tools with a variety of academic criteria.

In nine studies correlating the TOEFL with GPA, the predictive validity coefficients ranged from .17 to .43. In two studies which used TOEFL and GRE (Graduate Record Examination) scores together to predict GPA, the highest multiple correlation coefficient was .32. In a similar study which used a GRE retest in the final term as the criterion, the multiple correlation coefficient was .71. In a variety of studies using other ESL tests as predictors of GPA, number of university credits earned, or academic success, predictive validity coefficients ranged between -.05 and .70. Davies and Howatt (1983), reviewing the predictive validity of University of Edinburgh's English Language Battery (ELBA) with academic outcomes over ten years from 1973 to 1983 showed predictive validity in the region of .3. Criper and Davies (1986) demonstrated a predictive validity of .45 for the ELTS test against final outcome after one year. Thorndike and Hagen (1969) state that predictive validity of .3 will yield correct choices 60% of the time; .45 will yield correct choices 65% of the time (p. 172).

One reason for low predictive validity is the diagnostic use of test scores to guide placement into language remediation courses concurrent with the academic course (Hamp-Lyons, 1986b). Another reason, when using academic outcome as the criterion, may be that as there is a very high success rate in all academic courses, both in the USA and Britain, the criterion permits of little discrimination. It may be that the predictors are discriminating among testees with unnecessary fineness, and thus obscuring predictive validity. Thorndike and Hagen (op cit) discuss the greater merit of expectancy tables in terms of useable information provided, and these do seem to be an easily interpretable way of presenting predictive validity data.

2.2.2.2.3 Criterion validity: postscript

As a postscript to the discussion of criterion validity, both concurrent and predictive, it is important to remember that in measuring criterion

validity it is not necessary to know what the test is measuring, as long as whatever it is measuring is a good predictor of the criterion. As Palmer and Bachman (1980) point out, a problem with criterion validity is that a test may have criterion validity without one's knowing what it measures.

2.2.2.3. Content validity

"Content validity is established by an expert appraisal of the test content as a sample of the subject to be learned" (Davies, 1978: p 61). Palmer and Bachman (1980) describe it as "... basically a sampling process (which) requires a fairly complete description of the type of competence being tested" (p.2). Thorndike and Hagen (1969) tell us:

> *"To the extent that our objectives, which we have accepted as goals for the course, are represented in the test, the test is valid....It should be clear that ...content validity is important primarily for measures of achievement" (p. 164).*

Davies (op cit, p. 62), while agreeing with this, adds that proficiency tests also require content validity. In proficiency tests, the language needs to be fully represented and adequately sampled. Moller (1982) believes that the content of a proficiency test reflects the test writer's decisions about the universe of content to be sampled and his choice of sample: in this view, the test evaluator looking at the test for content validity is really assessing the constructor's definition of proficiency.

Carmines and Zeller (1979) describe three steps to achieve content validity: 1) specify the full domain of content; 2) sample appropriately from the domain; 3) put the sample into testable form (p. 21). Their view is that it is usually impossible to sample content; instead, a set of items which it is hoped reflect the content of a given (in theory) content domain are usually constructed. They remind us that there is no

agreed criterion to determine the extent to which a measure possesses content validity: rather, this rests on an appeal to reason. Cronbach (1971) adds that the content of a test should be judged for accuracy as well as for relevance to the universe. For this a subject area specialist is required. In addition, the test specifications need to state the characteristics of distractors, where these are used, because the content of an item can be altered by altering the distractors.

2.2.2.4. Construct validity

Anastasi (1976) says that "The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait" (p. 151). In simpler terms, Cronbach (1971) tells us "Every time an educator asks 'But what does the instrument really measure?', he is calling for information on construct validity" (p. 463). Jakobovitz (1970) is referring to construct validity when he says:

> *"The question of what it is to know a language is not well understood and consequently the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well defined". (p. 75)*

Carmines and Zeller (1979) see construct validity as central to the measurement of abstract theoretical constructs (such as language proficiency). Anastasi (1976) regrets the fact that some testers have presented construct validity as "...purely subjective accounts of what they believe (or hope) a test measures" (p. 160). Such subjective accounts are, in this view, merely a special kind of face validity - face validity to the language tester - until the construct in question can be described and observed. Anastasi believes that this perception of construct validity may have arisen from the practice of describing construct validity as theoretical, and from statements such as that of Cronbach and Meehl (1972): "...construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not

'operationally defined' " (p. 91). Such statements have been interpreted as suggesting that the construct is not amenable to proof.

Anastasi stresses empirical techniques for measuring the theoretical construct, e.g., correlations with other tests, factor analysis, internal consistency measures and convergent/discrimination by the multitrait-multimethod matrix. Anastasi (1982) says:

> *Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations are grist for this validity mill. (p.144)*

As long as empirical investigations of construct validity are not carried out, Morrow (1981) has some basis for claiming that construct validity is circular. Porter (1983) believes that construct validity need not be a circular concept if empirical evidence is required to show that the construct has some reality independent of other constructs, i.e., the theory of language (learning) upon which the test is based must be stated explicitly and supported empirically.

Carmines and Zeller (op cit) describe three steps in construct validation:

> *1) the theoretical relationship between the concepts must be specified;*
>
> *2) the empirical relationship between the attained measures of the concepts must be examined;*
>
> *3) the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure. (p. 21)*

In interpreting the empirical evidence, the test evaluator must be able to state several theoretically derived hypotheses involving the particular concept, to which support has been given by the study. Carmines and Zeller point out that construct validity is not established by confirming a single prediction on many occasions or by confirming many predictions

on one occasion. Ideally there should be a pattern of consistent findings involving different researchers using different theoretical structures in a number of different studies. In contrast, statements such as that by Cronbach and Meehl, that it is essential that construct validity is "...not identified solely by particular investigative procedures, but by the orientation of the investigator" (1972: p 97) tend to obscure the concept again, suggesting that construct validity should be subjectively rather than empirically determined.

Weir (1986) argues that there is an equally important need for construct validation at the *a priori* stage of test design. He believes that:

> *...the more fully we are able to describe the construct we are attempting to measure at the a priori stage the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test. (p.2)*

In this view, statistical analyses of test results would serve *a posteriori* confirmatory construct validation functions.

What appears to be needed is a combination of hypothesis formation, empirical investigation and hypothesis testing in rigorous contexts, replicated many times, that is, a combination of *a priori* and *a posteriori* approaches. In this view, the most satisfactory definition of construct validity to date seems to be that of Messick (1975):

> *"Construct validation is the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning." (p. 955)*

Construct validity has been receiving increasing attention in language testing in recent years, as evidenced by such studies as Oller and Perkins (1978), Bachman and Palmer (1982), Upshur and Homburg (1983),

Alderson and Urquhart (1983) and Klein-Braley (1985). Vollmer (1981) has said:

> *We as a society simply cannot afford to classify people and divide them up....as long as the question of construct validity of the instruments used is not clarified somewhat further. (p. 168-9)*

As Alderson (1981) suggests, there is potentially a conflict between construct validity and content validity. For example, the skills/aspects model may provide a convenient basis for sampling of items for tests, i.e., it may facilitate content validity. It may, however, prove not to reflect the true nature of how language functions, i.e., it may have poor construct validity. Alderson sees this as a question of whether tests are "...mirrors of reality, or constructed instruments from a theory of what language is, what language processing and producing are, and what language learning is".

2.2.2.5. Reliability-validity : tension?

Davies (1978) suggests that:

> *"In testing as in teaching there is a tension between the analytical on the one hand and the integrative on the other... The two poles of analysis and integration are similar to (and may be closely related to) the concepts of reliability and validity. Test reliability is increased by adding to the stock of discrete items in a test... Validity, however, is increased by making the test truer to life, in this case more like language in use". (p.49)*

It is from these and similar comments by Davies that the now frequently-accepted view of 'reliability-validity tension' has arisen. Since both of these expectations of language tests are fundamental and critical, this point must be investigated.

Thorndike and Hagen (1969) tell us: "Validity, insofar as we can appraise it, is the crucial test of a measurement procedure. Reliability is

important only as a necessary condition for a measure to have validity" (p. 189). Lado (1961) makes the same point: "Reliability is necessary for validity, because a test with scores that fluctuate very much does not test anything" (p. 31). Spearman (1936) made this clear many years ago:

> *"...the interrelations of reliability and validity are one sided. Low reliability necessarily means low validity, but the converse is not necessarily true. Whenever we find bad agreement between different measurements, then we can safely say that the examination is bad. But when the measurements agree we can not forthwith say that the examination is good". (p.108)*

It must be remembered that test reliability refers only to the proportion of the variance observed in a set of measurements that is true variance rather than error variance. Spearman's statement above indicates that when two (or more) measurements using the same test disagree very widely, the explanation lies in poor content or construct validity. But when these two measurements correspond, we can say only that whatever the test measures, it appears to be doing it reliably. It might in fact be testing something other than what was intended: Oller (1978) and Gunnarson (1978) discuss this point in detail. Davies' (1977) statement that "Reliability is most happily seen as a form of validity" (p. 38) thus becomes simpler to undertand and to accept: without reliability there is little point in considering a test's validity, since it is not possible to identify what the test is measuring in order to consider any of the theoretical validities. Where then, is the reliability-validity tension? Only, it would seem, in the degree of reliability which is demanded before a test can be investigated for validity.

Lado (1961) pointed out that we often "have to choose between more apparent validity but less objectivity, and more objectivity but less apparent validity" (p. 29). Guilford and Fruchter (1978) develop this point in detail. They focus on the mathematical incompatibility of reliability, which requires high item intercorrelations, and validity,

which for many purposes requires items which are of varying difficulties and which measure different factors: such items would necessarily have low item intercorrelations (assuming each item to be valid). They make two suggestions: first, a degree of compromise. They believe that inter-item correlations for well-constructed items which range between .10 and .60 will yield reliabilities approaching .90. The other solution would be to use a battery of tests. For each test the goal should be reliability, though some reliability should be sacrificed for the sake of a difficulty range of items. Each test should be designed to measure one common factor, which means there should be minimal intercorrelations between tests. The goal for the test battery should be validity.

This discussion brings us back again to the issue of reliability - validity tension: if such a tension indeed exists, it is because we have insufficient knowledge of the nature of language, of language learning, and of language use, so that our perceptions of these constructs differ, and therefore we do not all interpret 'validity' similarly. Spearman had this to say in 1936:

> *"...that which is intended to be measured by examinations is generally most vague and equivocal...Save where there is some sort of answer available to this great question as to purpose (of examinations), examinations...would seem to be but a groping in the dark." (p. 109)*

As language testing comes closer to defining what must be tested, why, and in what ways, i.e., to greater content, criterion and construct validity, relationships between measures will become easier to establish *a priori* and the improved test design will result in fewer problems of reliability, and the tension will dissipate. That position is, however, still in the future.

2.2.3 Practicality

The concept of the practicality of a language test was introduced by Lado (1961). Developing the point, Harris (1969) lays down some aspects of practicality which must be considered:

1) *economy: copy cost; cost of scoring time (human or machine); administration time*

2) *ease of administration: equipment; space; number of administrators*

3) *ease of interpretation of scores*

To this list we may add the level of qualification required of administrators and scorers. Corder (1973) points out that as testing techniques become more sophisticated, they also become more expensive. However, Hughes (1981) argues that valid tests may save money, by avoiding wrong decisions.

Cronbach (1971) describes a utility equation which can express the practicality of the use of a test for selection decisions. The equation incorporates information on the importance of the decision, the cost of testing, the proportion of applicants who will be accepted, and the predictive validity coefficient. Cronbach does not think that the equation is likely to be applied formally by test developers, but nevertheless it is conceptually useful. Essentially, the cost of testing must be weighed against improved accuracy of decision-making. Cronbach also reminds us that in many cases a testee's scores are placed on file and used for a number of later decisions: each use should be offset against the initial cost of testing.

2.2.4. Backwash

Backwash ("the receding effect of a wave, literally and figuratively": Concise Oxford Dictionary) refers to the effects of testing on teaching. Heaton (1975) says:

> *"It can be argued with some justification that language examinations in the past have exerted a harmful influence on the language teacher and have considerably inhibited language learning...." (p. 21)*

Jackson (1965) takes an extreme position on the same point:

> *"All examinations in English purport to be testing devices. All experience shows that their techniques immediately become teaching devices." (p. 13)*

Carroll (1973) agrees that "this matter of the relation between in truction and examining is certainly one of the persistent problems of foreign language testing" (p. 16), but points out that, while examinations can have adverse effects on teaching, they can also have beneficial effects. He sees the solution to 'teaching for the test' as being to make better tests. Jones (1977), making the same point about the backwash effect of testing, suggests that students are "more concerned about immediate realities such as tests and grades than they are about any stated objectives for which they will not be held accountable" (p. 241). He also suggests that teachers feel the pressure to teach to the tests most strongly when several colleagues have students taking the same test. Heaton (op cit) considers it fair to point out that testing has been "one of the greatest single beneficial forces in changing the direction of language teaching in many areas" (p. 162). Both Jones (op cit) and Carroll (op cit) highlight the introduction of listening and speaking tests as having had a beneficial backwash effect in increasing the amount of attention paid to oral production and listening comprehension in EFL instruction. Davies (1977), in discussing a test of Spoken English in West Africa, saw one of the demands to be made on the test to be that

its effects should be beneficial, in terms of encouraging the teaching of spoken English and acting as a goal for that subject, and also in terms of raising the standard of spoken English.

In designing new test instruments, backwash effect must always be a consideration, and the expectation will be that the new test will have a more beneficial effect, or at least not less so, on teaching.

2.3. Norm-referenced and criterion-referenced tests

In educational measurement there are essentially two different kinds of frames of reference for assigning and interpreting test scores. In norm-referenced procedures, information from a group of individuals provides a frame of reference against which the score of each individual can be compared. In criterion-referenced procedures, the test provides its own frame of reference.

2.3.1. Norm-referenced tests

Norm-referenced tests provide information by comparing the scores of each individual to those of other individuals tested by the same procedure. This is done through the application of more or less sophisticated measurements such as ranks, percentiles, grades and standardised scores. In each case, the basic principle is the same: the measures provide a continuum from best to worst, and each individual is located at a fixed point on that continuum (Pilliner, 1978). In a norm-referenced test, success is defined in terms of the performance of all the candidates taking the test, and thus can shift from occasion to occasion according to the distribution of performance within each norming group. Brown (1980) points out that with norm-referenced tests we may have some general knowledge of what the assessment measure was concerned with, but that normative scores tell us only whether the testee knows

more or less than other testees, not what it is that testees know or do not know, can or can not do.

Norm-referenced tests have been the standard kind of test since they were introduced early this century, as part of the response to mass education and the need for efficient selection procedures for further education and training. The virtues claimed for norm referencing are objectivity, stability and comparability: essentially these are all reliability claims, although there are also practicality claims for norm-referenced tests. The basic principle underlying norm-referenced tests is that if you have a large enough sample all human performance falls into a 'normal' distribution from extremely good to extremely poor, with the majority around the average or 'norm'. A norm-referenced procedure is designed to produce a bell-shaped curve representing the distribution of individual scores around the mean. To produce this distribution, a norm-referenced test must contain some very easy items that all testees can perform correctly, and some very hard items that no testees can perform correctly. Test items are constructed, selected and rejected on this basis. Clearly, the ability of the population against which the test is normed has a great effect on the shape of the curve and the point on that curve at which individuals will appear.

Since the information produced by a norm-referenced test is relative, these tests are most appropriately used in situations where comparisons among testees are sought, such as in a competition for a limited number of scholarships for higher education courses. Brown (op cit) points out that norm referencing is basically an aptitude procedure, since any items which all testees get right (because thay have achieved a certain level of performance) will be rejected. Cresswell and Houston (1983), however, believe that the situation is not that clear, and that norm-referenced and criterion-referenced tests become harder to distinguish once one begins to consider how standards (norms) for performance are set.

2.3.2. Criterion-referenced tests

Brown (1980) defines criterion referencing in general terms as follows:

> *Assessment that provides information about the specific knowledge and abilities of (pupils) through their performances on various kinds of tasks that are interpretable in terms of what (they) know or can do, without reference to the performance of others. (p. vii)*

A criterion-referenced test includes the specification of a level, or several levels, of previously defined and described competence, and each individual's performance is judged in terms of how it measures up to the specified level(s) of competence. It follows that on any one testing occasion the number of testees at any particular level of performance can shift dramatically from any other testing occasion where the same pre-specified levels are applied, according to the chance distribution of abilities of the testee group on that occasion. Because they are based on test performance criteria rather on the performance of a "normal" population, criterion-referenced tests do not seek to achieve a normal curve. Those who know and can do what is necessary for this test will score high while those who do not will score low. Items for a criterion-referenced test are chosen not for discriminatory power but to represent a range of relevant tasks. Scores on criterion-referenced tests provide information not about testees' relative standings but about what each one has and has not achieved in a particular area of study or skill. Thus, criterion referenced tests have great potential as diagnostic instruments.

It can be seen that the information from a criterion-referenced test is more specific than that from a norm-referenced test. With a norm-referenced test, performance on all the items is usually summed and reported as a total global score. In contrast, a criterion-referenced test is designed so that inferences can be drawn from the testee's performance on each of the test items, and the global score would be irrelevant. Pilliner (1978) points out that it may happen that different

testees have quite different patterns of item scores which add to up similar or identical global scores: a norm-referenced test would take no account of this possibility. Pilliner describes these 'profiles' as "of the essence" for a criterion-referenced test designed to tell whether each individual reaches a previously-specified level of acceptability on each part of the test (p. 39).

The virtues claimed for criterion-referenced tests centre on validity and on the increased possibility of positive influence on instruction. The problems relate to the difficulty of transferring a criterion-referenced test from one context, for which it was expressly constructed, to another; and to the difficulty of arriving at satisfactory statistical measures for describing and generalising test performances. Classical test statistics are of little applicability, since these tests are not designed to discriminate among testees, and discrimination is at the heart of classical reliability measures. Recently work has centred on estimation of the standard error of measurement, and the application of Item Response Theoretic approaches, most usually Rasch analysis.

Cresswell and Houston (1983) argue that decisions about what to include in a test and about criteria for performance levels are made with reference to experience of what a reasonable proportion of testees are likely to be able to achieve. This is confirmed by Pilliner's (op cit) description of procedures for setting 'cut-off' scores for criterion-referenced tests (pp. 44-45). In some sense, then, a criterion-referenced test is always a norm-referenced test.

2.4. Concepts in language testing: implications for testing writing in academic settings

We may derive certain basic principles from the preceding discussion which must be applied to an investigation of any writing test, whether general purpose, academic purpose or specific academic purpose.

While reliability is not the most important test characteristic, unless the writing test is reliable its other characteristics cannot be meaningfully investigated. For a writing test which is used operationally to make significant decisions about whether or not testees are awarded scholarships, or are permitted to participate in career-advancing courses of study, we hold the view that a high degree of reliability is required - at least .80.

Reliability is necessary but not sufficient for any test: it is essential to look carefully at validity also. Construct validity is the most difficult validity to establish, but it is also the validity a test of writing in an acdemic setting, whether that setting is general or specific, needs most. We hold the view that *a priori* and *a posteriori* construct validation are both necessary; this view has motivated the first section of this chapter, in which the attempt was made to arrive at a construct of writing, with particular reference to writing in a second language. The same view motivates the exploration in Chapter 3 of the constructs of language proficiency, English for Specific Purposes and English for Academic Purposes.

For an operational writing test, such as is the focus of this study, considerations of practicality and backwash are also important. Direct tests of writing are less practical than objective measures which correlate highly with them; on the other hand, direct writing tests are generally claimed to have beneficial backwash while objective measures are claimed to have negative backwash. The practicality and backwash of dirct writing assessment are discussed in Chapter 2.

In considering whether writing tests should be referenced against a fixed set of criteria, with each testee being measured on the criteria without comparison to any population, or referenced against the performance of a large population, with each testee being measured by comparison with the other performances on the continuum and then being placed at an

appropriate point on that continuum, the problems of reliability-validity tension surface again. White (1985) sees the problems with using norm-referenced writing tests as located in test design and in the norming populations. If the test measures what it claims to measure (i.e., is valid), and if the norming population is an appropriate and sufficient one for the target testee population, he believes that norm-referenced writing tests are acceptable. If these two demands are not met, however, he believes consideration should be given to a writing test linked to pre-specified standards and not to a population (p. 67). In this case, the test design process will entail the specification of what the test is intended to measure, how it will do so, and how satisfactory performance at the performance levels can be recognised, that is, it will ensure maximum validity.

White (op cit) cautions that there is a danger in criterion-referenced testing of writing that unrealistic or unfair criteria may be applied, or that criteria which were fair and realistic in one context may be transferred into another context for which they are inappropriate. In other words, it is easy for poorly designed or inappropriately applied criterion-referenced tests to be invalid. White prefers a procedure which blends norm referencing and criterion referencing, in which a scoring guide sets out criteria for scoring while papers are ranked against sample papers representing points on the scoring continuum. With careful monitoring, such a procedure combines attention to validity with attention to reliability. Procedures for the scoring of writing tests, together with a detailed discussion of all the expectations of and variables in writing assessment, are the focus of Chapter 2.

It will be difficult for any writing test to meet the expectations suggested in this survey: nevertheless, the attempt is essential, as is the reporting of information concerning the extent to which the test has succeeded in fulfilling them.

CHAPTER TWO

THE PROBLEM IN CONTEXT (1): THEORY AND PRACTICE IN WRITING ASSESSMENT

INTRODUCTION

The use of direct writing samples as the basis for judgements has been common in Britain and the United States of America since the Victorian era, and has been problematic throughout that period. Writing tests which require a sample of writing from the testee are often claimed to possess inherent validity, on the basis that the skill, or proficiency, being tested is the same as that being used to answer the test. However, they have long been considered inefficient and unreliable, in contrast to multiple-choice tests, many of which correlate satisfactorily with direct measures. Traditionally in testing, indirect measures are preferred over direct measures only when the indirect measures show clear advantages, and this was the argument behind a great deal of direct assessment of writing in the past forty years. We have recently emerged from a period in which proponents of indirect measures could point to the high cost and low reliability of direct tests of writing and make strong claims for the greater efficiency of indirect measures. This battle has been fought and won on the issue of backwash, especially in the U.S.A. (section 2.3.).

Many problems remain unsolved or only partially solved, however, and these are explored in the sections which follow. Four key variables in the construction of a writing test: the task; the writer; the scoring procedure; the rater, are considered in detail, in reverse order. The chapter ends with a consideration of how research in the assessment of writing, the great majority of which has been the assessment of L1 writing, can be applied to the assessment of writing in the second language, and in academic settings.

CHAPTER TWO

1. WRITING TESTS

1.1. Definition

In the ensuing study, and in this chapter, the term 'writing test' refers to assessment based on a direct sample of the testee's writing. There are two kinds of test which, within the definition above, may be considered as 'writing tests'. One is a test of anything which uses continuous writing as the test method. The other is a test of writing which uses continuous writing as the test method. The focus of the study which follows, and of this chapter, is on the second, although as we shall see in the next section the second grew out of the first, and as we shall see when we turn to the research in Chapters 4 to 6, in testing writing in academic settings the distinctions between the two become blurred.

In our terms, then, a writing test is the collection of a direct writing sample or samples from candidate writers in a highly structured test context, in response to an assigned task, which sample(s) is/are then read and rated by one or more qualified judges (raters/readers) using a procedure which has been more or less closely specified and is more or less replicable by other readers.

1.2. History

Cox (1966) describes the establishment of traditional examinations as a relatively late development from the "rather haphazard medieval system" (p.2) and resulting from the advent of mass education and increased public expenditure on education. Brooks (1984) attributes the development of public examinations to a social change from patronage to proven ability, to the competitiveness of the age of imperialism and to the increase in size of the middle classes. Many of these causes are the

same as or very similar to the causes of the rise in literacy described in chapter 1, section 1, and clearly there is a very strong relationship between the testing of writing, particularly academic writing, and the spread of mass literacy. In the period from 1800 when the Oxford public exams statute was passed to the Schools Enquiry Commisssion of 1868, there was an enormous increase in the number and specificity of public examinations, to the extent that the Report of the 1868 Commission complained that the qualification bandwagon was in danger of making "effective organisation of the school as a place of general education impossible" (Vol.1, p.324: quoted in Brooks, op cit). For many years the traditional essay-type examination which was the method used for all these public examinations was apparently considered to be a perfectly adequate instrument. What little criticism there was focussed on the lack of consistency of these examinations, and there appeared to be little concern that the examining of such subjects as mathematics, physics and chemistry by the essay method might be invalid.

Attention began to be focussed on standards within and between examinations, and a small number of reports such as Edgeworth (1888; 1890) and Starch and Elliott (1912) showed that there were considerable differences between examiners. By early this century there was clear concern that standards were not fixed, and it was gradually realised that the test method, or more accurately the scoring procedure, was at the heart of the matter. Ballard (1923) said:

> *One of the defects of the essay as a measuring device...is the impossibility of making (it) amenable to rigid objective measurement...An essay is an intricate mental product which cannot be analysed completely...it fails through its very wealth and complexity (p.61-62)*

Concerns about the inconsistency of written examinations were at a peak in the 1920s and 1930s: a number of attempts had been made to improve the consistency of scoring of essays. Britton et al (1966) describe work by Rice (1903) and Hillegas (1912); Willing (1918; 1926) divided the

composition into 'style' and 'form', thus introducing an analytic element; Val Wagenen (1920) analyzed the composition into thought, content, structure and mechanics, and provided criteria for each. Although there were some encouraging reports (Willing, op cit; Van Wagenen, op cit; Hudelson, 1925; Sims, 1933; Stalnaker & Stalnaker, 1934; Traxler & Anderson, 1935), there was also a good deal of contrary evidence (Hulten, 1925; Valentine, 1932; Hartog & Rhodes, 1935; Hawkes, 1936, Stalnaker, 1937; Cast, 1939,1940).

Two major reports, one in Britain and one in the United States, were primarily responsible for a dramatic shift away from essay examinations and toward standardized, or "objective", testing in the late 1940s and 1950s. The first, by the International Institute Examinations Enquiry Committee, under the chairmanship of Sir Philip Hartog, reported a series of studies begun in 1935 which were published in 1941 as The Marking of English Essays. Their study showed serious inconsistencies in marks awarded, and little improvement as a result of various innovations they attempted, and was then, and is still generally, understood to be an indictment of the English composition test. However, careful study of the Report shows that Hartog et al were oversimplified and misunderstood: their criticisms were directed at poor testing practice, such as the "vast or vague" subjects set for essays (p.138), rather than at the test type itself. Their first recommendation was:

> *That the practice of asking pupils from the age of 13 and upwards to write "compositions" termed "essays" be abandoned; and that they be asked instead to write compositions on subjects about which they may reasonably be expected to have a fund of ideas and a sufficient knowledge which they could express for a given audience and with a given object in view (p.142)*

As we shall see later in this Chapter, these recommendations are very close to the "discoveries" made in the last ten years about good writing test practice.

The second report was by the College Entrance Examination Board in 1946, which showed reader reliabilities of .55 and concluded that "the problems involved in developing a reliable essay examination are, if not unsolvable, at least far from solved at the present time." (quoted in Huddleston, 1954: 166). An earlier report by the same body, in 1931, had called for the retention of essay exams despite their limited ability to predict academic performance, giving as the reason the importance of ensuring that American culture would retain values of civilization and culture rather than of mechanical efficiency. By 1946, the pressure had become too great, and the conclusions of this Report, and in-depth evaluation of CEEB data, led to the replacement of the essay test with a standardized test of "verbal ability".

If the testing of writing by writing fell into disfavour with measurement experts, it nevertheless remained popular with English teachers, who viewed standardized testing with suspicion. Wiseman (1958) was one of the strongest opponents of standardized testing of writing, recognising the possible educational consequences well before these began to appear. The essay exam never disappeared from the British education systems as it did in the U.S.A. Most educational research in that country in the 1950s and 1960s was focussed on the improvement of standardised testing, although a few lone voices, such as Stalnaker (1951) expressed concern about the effects of standardised testing on literacy in the school. Stalnaker believed that objective tests could not tap higher order mental functions, while essay tests can stimulate good study behaviour and encourage students to see writing as a means of of learning.

In British schools and colleges, the testing of academic subjects in schools and colleges by writing has continued undaunted by the findings of reliability studies. Education systems, and particularly higher education systems, have introduced a wide range of other test methods but have continued to use direct writing tests as well. Britain has not suffered the crisis of literacy which was felt in the U.S.A. in the late

1970s when it was discovered that standards of literacy in schools and colleges had declined precipitiously since the 1960s (Cohen and Brawer, 1983). Bishop (1978) placed some of the blame for this on the decrease in attention to writing in schools, which could in turn be blamed on the decline in assessment through writing in American educational institutions. In addition, standards of educational achievement also declined dramatically during this period, and although other reasons also exist, Bishop states a common view when he links this decline also to the decrease in the amount of writing done in schools. In the U.S.A., direct writing tests have gradually found their way back into the battery of methods used for English testing (whether L1 or L2), culminating this year with the introduction by Educational Testing Service of the 'New TOEFL Writing Test'.

2. TESTING WRITING: PRINCIPLES

2.1. Writing tests: validity

What makes the direct testing of writing valid? We may begin with the most obvious argument, put forward by many teachers and researchers, among them Diederich (1974):

> *As a test of writing ability, no test is as convincing to teachers of English, to teachers in other departments, to prospective employers, and to the public as actual samples of each student's writing... (p.1)*

The argument is, of course, one of face validity. Jacobs et al (1981) make the same point when in their list of arguments in favour of direct writing tests for the assessment of the writing performance of L2 learners they say: "a direct test of writing is an unarguably valid measure of writing proficiency". (p.3)

2.1.1. Beyond face validity

The face validity argument put forward by Coffman (1971) is of a somewhat different kind: "...the persistence of essay exams reflects the judgement of teachers that no effective alternatives are available." (p.24) Thus, despite the traditional and well-known problems of writing tests which are discussed below (Section 2), writing tests appear to possess certain qualities which for many teachers override their disadvantages. Breland and Gaynor (1979) suggest one of these qualities when referring to face validity:

> *Indirect measures lack face validity and credibility among members of the English profession and educators generally, and they tend to deliver a message to students that writing is not important. (p.127)*

In other words, the fact that writing is directly tested gives it face validity in the eyes of the students . Coffman (op cit) argues that writing tests have content validity: "The essay examination constitutes a sample of scholarly performance; hence, it provides a direct measure of educational achievement." (p.24) In the responses to the questionnaires which were collected from academic supervisers as part of the University of Edinburgh/Institute for Applied Language Studies' ELTS Validation Project (Criper and Davies, 1986), comments from supervisors frequently supported Coffman's argument: asked whether there was any other information about an applicant's English language proficiency they would like (apart from ELTS test scores), a number of supervisors replied that they would like to receive a sample of the applicant's writing. A similar response was received from academic faculty surveyed by Bridgeman and Carlson (1983) when they explored possible writing test design in preparation for development of the TOEFL Writing Test. In a small survey of faculty at the University of Edinburgh who teach large numbers of overseas postgraduate non-English speaking students by this researcher,12 ofthe 24 respondents said that they based 75% or more of their final coursework grade on written tests.

2.1.2. Invalidity?

In contrast, Moller (1982) criticised the direct testing of writing because it is examiner-based rather than language-based; because the evaluation criteria are unclear; because assessment is subjective; and because scores awarded are based on precedent and the assessor's knowledge/experience. The four points are essentially one objection: the direct evaluation of writing is centered in the human evaluator, and therefore the possibility of human error exists. That Moller's objection is not to the testing but to the scoring procedure is made clear by Pilliner's (1968) lucid statement of the difference between subjective and objective testing. Pilliner points out that of the three processes common to all examining, (a) the construction of questions, is clearly subjective, requiring judgements such as selection from within a domain, choice of topics, item priorities and framing of questions; (b) answering the questions, is also clearly subjective even when the question format is 'objective' because the testee must make choices based on personal judgements; only (c), scoring or marking, can be divided into both subjective and objective possibilities: if marking is done by a machine or machine-like process, scoring is objective (i.e., there is no room for decisions). Nevertheless, the assignment of the 'right' answer on an objective test is another subjective process requiring decisions by a human test constructor. Moller's criticisms are, then, directed at the scoring methods rather than at the construct, at the reliability rather than the validity aspects of writing tests. These criticisms are dealt with in detail in the next section.

2.1.3. Construct validity

The argument put forward by Braddock, Lloyd-Jones and Schoer (1963) in favour of the direct testing of writing is one of construct validity:

> *Not only do they not require the examinee to perform the behavior being measured - he does not do actual writing -*

> *but these tests also make no attempt to measure the "larger elements" of composition even indirectly. (p.42)*

Jacobs et al (1981) also stress the construct validity of direct writing tests:

> *"the direct testing of writing emphasizes the communicative purpose of writing" ...a direct test of writing "utilizes the important intuitive, albeit subjective, resources of other participants in the communicative process - the readers of written discourse, who must be the ultimate judges of the success or failure of the writer's communicative efforts." (p.3)*

Jacobs et al put forward another argument in favour of the direct testing of writing which relates to the discussion of proficiency in the next Chapter. They believe that a composing task "invokes and challenges" the writer's general language proficiency, and may therefore be a good measure of general language proficiency and not only of writing proficiency. In addition, a writing task can be expected to activate specific proficiency factors as well as the global factor depending on variables within the task (p.4).

Having resisted the introduction of direct tests of writing for many years, TOEFL have this year (1986) instituted their New TOEFL Writing Test and explain their decision primarily on validity grounds:

> *...researchers and educators have begun to modify the definition of writing competence... Angelis reported the perception among graduate faculty that there may be little actual relationship between recognition of correct written expression and the production of an organized essay or report. Direct measures of writing, such as essay tests, are increasingly viewed as being a more valid approach to writing assessment. (Stansfield and Webster, p.1)*

Jacobs et al (op cit) had previously suggested that although writing test scores can correlate highly with indirect measures of writing such as the Writing Ability section of the TOEFL (Pitcher and Ra, 1967) overseas students with satisfactory TOEFL "writing" scores often find the

university-level writing they must do a serious stumbling block, and are often assigned to special or "remedial" writing courses. This is a situation familiar to teachers on EAP pre-sessional or concurrent programmes at British and American universities. Jacobs et al suggest that this means there is some information about their actual writing ability which is not being obtained by these indirect writing tests. Burgess and Greis (1970) studied the performance of ESL freshmen on a range of language proficiency tests against Grade Point Average and Cumulative Grade Point Average as criterion variables, and found that the best predictor of both was writing grade in ESL Freshman English Writing. Shohamy and Reves (1985) remind us that, although a number of studies in the 1970s found high correlations between direct and indirect tests claimed to be testing the same skill or trait, correlational data cannot be interpreted as proving that the two measures were measuring the same thing. They cite several studies which show that the method of testing affects the assessment of the trait, and interpret these studies as showing that it cannot necesarily be claimed that direct and indirect tests are "the same" (p.50). Underhill (1982) also points out that high correlations between scores on direct tests of writing and on indirect tests of writing must be treated with caution:

> *...the interpretation of correlations is far more complex (and subjective!) than the correlations themselves. Especially in the field of language testing, a person's interpretation of a set of statistics may depend entirely on the assumptions of his particular theoretical viewpoint: the statistics have no inherent meaning other than as a purely mathematical relationship between two sets of numbers. (p.32)*

A collection of papers (Language Testing 2,1) explored authenticity in language testing, and focussed on oral testing, but much of the discussion is relevant to a consideration of the validity of the direct testing of writing. In the collection, Stevenson criticises what he sees as a current tendency to side-step validation principles and procedures, in particular, to make "the perilous jump from simple to abstract, from face

validity to construct validity: we think it's valid, therefore it is!" (p.46: Stevenson's emphases). Spolsky's paper in the same collection takes a very different view, criticising " 'hocus-pocus' scientists" who "sometimes even claim not to care what they measure provided that their measurement predicts the criterion variable" (p.33). It will remain difficult to settle this argument as long as we remain without formulated and validated constructs of language proficiency in general and of writing proficiency in particular, as Raatz suggests in the same collection:

> *Construct validation is the approach usually chosen if there is already a theory available which refers to the trait to be measured. It is problematic if a theory has to be specifically developed for the test, because then we get a vicious circle where the theory is validated by the test and the test is validated by the theory. (p 62)*

As suggested by Weir (1986) *a priori* construct validation is the remedy for this.

2.1.4. Specifying a construct for writing tests

The study of research in writing and composing in Chapter 1 was an attempt to formulate a construct of writing/composing behaviour in general, and in an L2 in particular. Although as we saw, the limited amount of L2 research so far places limitations on the confidence with which a construct of composing in a second language can be viewed, we were able to draw from that chapter some basic features of the construct which can be applied in assessing the validity of writing tests.

If writing is developmental, tests of writing should encourage and guide development through the values they convey. Judgements made on the products of writing tests should take into account where the writer is in her personal growth as a writer, and how the specific task of the writing test relates to her current stage of development. Multiple samples based

on tasks at different levels will provide a more informative picture of the upper and lower limits of this than any single sample.

If composing is a heuristic procedure, a voyage of discovery, then what is discovered is both how to write and what to write. A construct valid writing test will present writers with a task which sends them on a journey to find out what they think and know in this area, at the same time providing them with a map to find their way, yet leaving the milestones blank for them to write their own directions. Reading, and other experience, can play an important part in this. Not all writers will travel the same route nor arrive at the same place. At the same time, the scoring method must make it possible to know the route the writer took and how far she travelled; it must also take into account the limits placed on the writer by the task and the test context, and the writer must not be penalised for the test developer's failure to appropriately operationalise the construct of writing in the test.

If composing is recursive and purposive, a writing test must provide the writer with an authentic reason for writing; it must also provide sufficient space for the composing processes to operate. The scoring procedure must take into account the extent to which the writer declared and fulfilled her purpose; and the extent to which she was constrained by the context of the test.

If writing involves meaningful communication with a reader, the writer needs an audience for her writing; and assessment by actual readers is essential. Readers need to give an authentic response while holding in the front of their minds all the constraints and limits on authenticity within which the writer wrote.

If writing is normative, the writer should know what norms are being applied and what values are placed on them; if the writer is informed of

the normative function of the assessment, she should be assessed on her approximation to those norms.

If composing is always a personal act and statement, even in extreme interpersonal contexts, the writer should be given space to make a personal commitment to her composed text and should be valued for such commitment.

2.2. 'Writing tests: practicality

Jacobs et al (1981) remind us that writing tests are relatively practical because an essay question can be set by any teacher, whereas the preparation of multiple-choice tests requires a great deal of special knowledge. A word of caution is necessary here, however: it is becoming clear that a valid and reliable writing test involves a considerable investment of time, money and expertise, as indeed Jacobs et al's own research shows. As the discussion in section 3.4 of this chapter shows, task variables are complex and presently little understood, and evidence is increasing that tasks have a strong influence on writers' performances.

Writing tests are practical to administer, requiring little typing and reproduction: for the same reason, they are quite cost-effective. Although the scoring of writing tests is labour-intensive, in many contexts teachers, who are accustomed to reading, responding to and grading many samples of student writing every term, are readily available to do this. To achieve an adequate level of reliability a commitment of time and money to developing scoring procedures, training readers and monitoring to ensure that judgements remain reliable is essential, but it is not yet clear that the commitment to an efficient system would be greater than the commitment to a similarly efficient system of standardised testing. The writing samples collected in the course of a writing test can be used for more than one purpose: they can be used to

provide personal individual feedback to the writer; they can be scored in such a way that they provide diagnostic information to assist in making appropriate placement decisions. They can be placed on file to form a developmental record of the writer over an extended period of time. The fact that most writing tests are not used in this multi-faceted way is regrettable but does not invalidate their potential practicality.

Finally, and perhaps most importantly, writing tests are quite practical to interpret: few teachers, parents or specialist score consumers find writing test grades (A-F) or scores (20-1; 9-0, etc.) difficult to interpret, and the provision of a simple explanation of the scores solves any such difficulties. It would appear that because score consumers think of an essay as a 'real' task, and one with which they have had direct experience, they are confident of their interpretations of scores.

2.3. Writing Tests: Backwash

While the value of literacy skills has never been questioned in developed cultures, they have come under subtle attack through the emphasis on measures of educational attainment which are 'objective', and therefore supposedly precise. We saw in the section on the history of writing assessment that the fears of Stalnaker and of Wiseman were well-founded. Teachers found that objective tests measured different qualities than subjective written tests, and found that to protect their students they had to teach to those tests. The backwash from standardised testing proved to be seriously damaging to educational standards throughout the U.S.A., and that country is still recovering.

In contrast, writing tests are generally accepted as having positive backwash: the pressure on teachers to 'teach to the test' in this case leads to more writing, more thinking and talking about writing in classrooms, more reading as input into writing, and a making overt the nature of reading and writing, of literacy, as power in the society

(Robinson, 1983). If it is believed, as Freire (1970) believes, that literacy is power, then writing skills cannot be neglected.

At the college and university level writing plays such a large role in success that every means of ensuring backwash into increased attention to writing skills in the pre-university period should be taken. The backwash effect of writing tests is particularly significant because, as is now believed (for example, Wilkinson, 1980; Jacobs, 1982) writing is a way of learning to write, and also a way of learning to think and feel. There are other ways, but the withdrawal of one of the most complete will leave an impact on not only the writing level but also the general educational level in the society. Many years ago, Meyer (1934, 1935) reported that practice in anticipation of an essay examination resulted in higher scores on other types of examination. An understanding of some these factors, coupled with an awareness of the importance of ensuring that students wishing to compete in modern society receive the right signals about what that society values educationally, prompted the introduction in 1985 of a direct writing test on the University of Cambridge's FCE (First Certificate in English): teachers of FCE preparation courses in countries around the world were concentrating on language structure, vocabulary in isolation and other such discrete tasks in an attempt to prepare candidates for the test (Foulkes, personal communication). The introduction of the optional writing test into the TOEFL, referred to earlier, was similarly motivated (Stansfield, personal communication).

2.4. Writing tests: reliability

It is generally accepted that reliability is a necessary precondition for any test, and it is on this point that most criticism of writing tests is based. Put briefly, the argument is that any test which is practical, has beneficial backwash, face validity, content validity and construct

validity but is unreliable cannot be valid, i.e., cannot yield meaningful results.

Godshalk et al (1966) distinguish two kinds of reliability which must be considered for writing tests: reading reliability and score reliability. Reading reliability refers to inter-reader reliability, and score reliability refers to the correlation between other essays by the same writers scored by different readers. Jacobs et al (1981) distinguish three types of reliability: score reliability, or writer reliability (i.e., degree of consistency of scores obtained by the same writer on different tasks and on different occasions); inter-reader reliability (i.e., degree of consistency of scores assigned to the same sample by different judges); intra-reader reliability (i.e., degree of consistency of scores assigned to the same sample by the same judge on different occasions).

2.4.1. Reader reliability

Reader reliability is the aspect of reliability normally referred to in statements about the unreliability of writing tests. Pilliner (1968) gives three reasons for reader unreliability: differences in severity level; variation in the range (spread) used; and rank order discrepancies. He suggests three approaches to improving consistency: the use of analytic marking schemes; increasing the number of questions; and increasing the number of raters. It should be noted that the second of these techniques does not apply only to reader reliability but also to score reliability, but as a reader reliability technique it has the same effect as his other two suggestions: that is, it is another way of arriving at multiple samples. Coffman (1971) identifies the same three causes of unreliability and makes a number of suggestions for improving reliability. First, he recommends that testees should identify themselves by numbers not names; second, that a sufficiently fine marking scale should be used (from 7 to 15 points is his recommendation); third, that model answers should be used to provide clear reference points to which

to tie the marking scale; fourth, that error should be distributed randomly not systematically by having more than one question and marking question-by-question, not student-by-student or by having more than one test; and finally, that multiple-marking should be used. These suggestions apply equally across and within readers, with slightly different strategies. Braddock et al (1963) recommend adopting a common set of criteria in preference to using model answers, but for the same purpose of reader reliability. Jacobs et al (1981) suggest seven steps for obtaining reliable readings:

> 1. *adopt a holistic evaluation approach;*
>
> 2. *establish criteria to focus readers' attention on significant aspects of the compositions;*
>
> 3. *set a common standard for judging the quality of the writing;*
>
> 4. *select readers from the same background;*
>
> 5. *train readers until they can achieve close agreement in their assessments of the same papers;*
>
> 6. *obtain at least two independent readings of each composition*
>
> 7. *monitor the readers periodically to check their consistency in applying the standards and criteria of the evaluation. (p. 28)*

We shall see in section 3.1 a wide range of reader reliability levels reported in the many studies to date. It is not possible to meaningfully compare reader reliability levels across studies because it is often not possible to discover what statistical procedures have been applied to arrive at the reader reliabilities reported. Ebel (1951) showed that the application of different statistical formulae led to markedly different reliabilities under certain conditions: because the term 'reliability coefficient' is a generic term it cannot tell us which method has been used. It would appear that the reliability levels reported in the early studies are single-reader reliabilities based on the Pearson product-

moment coefficient. Later studies are likely to have used more sophisticated procedures, either the Pearson product-moment coefficient with various corrections applied, or, as in the case of the Jacobs et al study, the Fisher intraclass formula as described by Ebel (op cit).

2.4.2. Score reliability

The principal method used for ensuring score reliability is multiple marking. Wiseman (1949) described a method, which became known as the 'Devon method', of using several markers marking by general impression and pooling their judgements, which provided the impetus for a number of methods of multiple-marking which have been developed since. Using four markers, Wiseman obtained reader reliabilities in the low .90s. This method called for rapid reading (50 scripts per hour of 30 minutes of writing by 11 year olds) but using markers known to be self-consistent.

Britton et al (1966) used the same method with three markers, who had not been trained or checked for self-consistency plus a fourth mark based on an analysis of mechanical accuracy, and obtained reader reliabilities in the high .70s. They take as their first principle that random error in measurement can be reduced by taking several readings and reporting the mean. They further believe that marker error should be reduced by taking the readings from different markers rather than taking multiple readings from the same marker, thus spreading the non-random error (i.e., any error which is inherent to the marker) randomly. They believe that when the measure being applied is itself guesswork these procedures are even more desirable. Pike (1973) similarly views multiple-marking as a method of combining fallible judgements to reduce random error, whereas Cox (1968) sees it as leading to a regression toward the mean and the washing out of whatever interesting and unique information a writing test yields over a standardized test.

In contrast, Wiseman (op cit) believes that some degree of marker variation is desirable, being an authentic diversity of response to complex and meaningful material: by combining the responses of several judges a composite mark would be arrived at which would be a more global view of the actual quality of the writing. Pilliner (1969) shows that if the markers used are highly self-consistent and at the same time agree poorly with each other, Cox's criticism would be justified. If on the other hand, there is a certain amount of agreement among the markers, the aggregate marks would be a valid expression of the amount of agreement between them. However, Pilliner also shows (personal communication) that although Wiseman is using the technique appropriately, the use of aggregate marks does not achieve what Wiseman claims: when aggregating occurs whatever is unique to each individual marker washes out, and the measure of agreement is an expression of common variance not unique variance. Finlayson (1951) found that using Thurstone's factor analytic method only one factor could be extracted from the marker behaviours; he feels that this shows that the markers were all seeing and valuing the same thing.

Swineford (1956) used only two readers with a third and occasionally a fourth moderating cases where the readers did not agree, and obtained a score reliability of .82 with trained readers. Pitcher and Ra (1967), using only two trained readers, obtained a reader reliability of .78. Jacobs et al (1981) report correlations for two readers, i.e., single reader reliability, from .59 to .96. When they used three readers the averaged reliabilities were between .89 and .94: thus three readers not only raised the reliability but made it more consistent. Studies by Follman & Anderson, 1967; Mullen, 1977; Flahive & Snow, 1980 also showed that the use of three readers consistently lifted reliability to the high .80s and .90s.

Although multiple marking is commonly associated with impression marking, it may be used in conjunction with any of the scoring methods

discussed in section 3.2, although it is least appropriate with the frequency count methods.

Multiple marking is a way of artificially creating multiple samples, thus increasing test length. Another way of improving score reliability is to increase test length, collecting multiple samples from each writer. Finlayson (1951) shows that scoring two essays by the same candidate and aggregating the marks has the same effect as using two markers instead of one, and similarly for three essays as three markers, and so on. In this case it is writer variability which is being washed out rather than rater variability.

2.4.3. Writer reliability

Writers write better or less well due to a number of factors, most of which are as yet poorly understood.

A number of obvious factors which will cause an individual's performance to vary from occasion to ocacasion on a writing test, as on any other test, have been well documented and all good testing practice takes them into account: time of day of test; physical environment; length of test; amount of time per task. However, even when all these factors are controlled, and when task variables are controlled also, we still find variation in writer performance from occasion to occasion (Kincaid, 1953). Writer variables will be discussed in Section 3.2, but because writing is a complex cognitive process, which engages the writer on so many levels simultaneously, it will never be possible to attain real writer reliability, since this would mean that the individual interaction between writer, material and task would have been washed out of the test instrument.

Problems of writer variability are usually accounted for by requiring multiple samples from the writer, which are either scored and reported

separately, or scored separately and averaged for score reporting. Vernon and Millican (1954) collected seven writing samples from trainee teachers over two weeks, and found a mean correlation between one writer's essays of .37 for the same reader, and .25 for different readers. These samples were not, however, collected in a testing context. Jacobs et al (1981) obtained a score reliability of .84 on two topics, with a correlation of .72 between mean scores on the two topics. Because the actual tasks are not given, it is not possible to know how much of this variation is due to writer variability and how much to task variability, i.e.,we do not know to what extent these were 'parallel tests'. The correlations across tasks intended to be parallel reported by Carlson et al (1985) are .71 and .68 respectively, giving score reliabilities of .83 and .85 with two raters.

Collecting multiple samples from each testee and aggregating scores results in higher score reliabilities. What is happening when this is done is the same as was discussed for multiple marking: the unique variance due to the interaction between the writer and the task is being washed out, leaving only the common variance of the writer's performance on all the tasks set. If enough tasks can be set to raise the aggregate score for the testee above .9, writer reliability will be statistically adequate. However, there is a decision to be made, whether the tasks set should be similar (in which case it will be easier to achieve high reliability levels) or should they be as different as possible, allowing the writer to show her full range of performance characteristics and increasing validity (in which case it will be more difficult to achieve high reliability levels).

2.4.4. Task reliability

We have seen above that even tasks designed to be parallel correlate only at about .7; in the Carlson et al study the 'parallel' tasks did not correlate more highly with each other than they did with tasks in a different mode. In both the Carlson et al study and the Jacobs et al study reader reliability, since it was an aggregate score, was high: above .9. If tasks were truly parallel we would expect then to correlate more highly than other tasks for which no such claim was made, or which were deliberately designed not to be parallel. No study has yet shown this to occur. This suggests either that writer variability is so great that it prevents task reliability from reaching reasonable levels, or that there are many task variables which have not been accounted for in the attempted design of 'parallel' tasks. The former may well be true, and not amenable to change, while the latter, if true, should in contrast be a situation which could be remedied.

2.4.5. Tension of expectations of writing tests

The reliability-validity tension identified by Davies (1978) appears to operate also in writing tests. Writing tests have a high degree of validity since the test method is the same as the channel for the trait tested; absolute validity cannot be claimed, however, as we shall see as the study in Chapters 4 to 6 progresses. A perfectly valid writing test needs to be valid on all possible variables, whereas the meaning of the validity claim usually made is only that writing is best tested by having testees write. In the ensuing study, we shall make much greater validity demands. It cannot be claimed that writing tests have ever achieved or will ever achieve perfect reliability, unless under totally impractical conditions such as the use of five or six raters. The use of human raters for writing tests makes this practically, although not theoretically, impossible. It is often claimed, in vindication, that the use of human raters necessarily makes a writing test valid, but this is

untrue. Ratings by human judges are not inherently more valid than ratings by mechanistic means: their validity depends on the extent to which they embody the construct which underlies the test and which it is intended by the test design should be measured. The same is true of machine scoring, since human judges made the decisions which the machine implements. If raters could achieve perfect agreement as to the constructs they were measuring and the ways in which those constructs are realised in human activity, there would be coincidence of validity and reliability. The reason human judges are used for the rating of writing tests is that we have yet to find any way of programming machines to recognise higher order mental functions or of judging such qualities as 'style' and 'voice'. Were we able to do so (and this is dependent on the development of the human skill, not on the machinery), we would be able to achieve perfectly reliable and perfectly valid assessment of writing.

There is also in writing tests a tension between practicality and backwash. Writing tests demand the investment of considerable amounts of the time of skilled personnel, often without recompense, or for unrealistically low compensation. Teachers rate written tests, rarely for whatever recompense may be available, and rarely with enthusiasm for the sacrifices the task involves, but because they have always been aware, consciously or otherwise, of the importance of what they do. Even the backwash from a poor writing test is far more positive than the backwash from a very reliable standardised test. In this tension, backwash has won over practicality consistently.

3. TESTING WRITING: VARIABLES

Braddock et al (1963) describe a number of variables which must be considered in any soundly based study into written composition: writer variable; assignment variable; and rater variable. Applebee and Brossell (1985) also identify three variables in writing assessment : topic variables; writer variables, and procedural variables. In this section, we shall combine these two systems, and look first at influences on the readers of compositions (reader variables); second at the influences of the actual scoring method used (procedural variables); next at factors affecting the writer (writer variables); and finally at features of the test environment with primary emphasis on the test question (task variables).

3.1. Reader Variables

'Reader variables' are those variables which affect the raters of student writing on writing tests: the term includes the features of the writing which (we believe) comprise the valid components of writing ability (i.e., the 'true' variance) and the features of the writing, the rater's perception of the writer, and the characteristics of the rater which cause different raters to respond in different ways to the same piece of writing (i.e., the 'error' variance).

3.1.1. What do readers respond to?

Probably the most detailed study of essay raters and their responses to various qualities in essays is Diederich et al (1961). The College Entrance Examination Board had spent six years developing a two hour essay test and training readers, in an attempt to achieve satisfactory reliabilities; however, they were consistently unable to achieve satisfactory composite score reliability, and thus Diederich et al were led to attempt to find differing 'schools of thought' among readers which

could account for variability in grading. Sixty readers were nominated by faculty at Educational Testing Services as being in high positions in their own fields and also concerned about writing: of the sixty, fifty three readers completed the assignment. None of the readers had previous training or experience in this type of marking, nor did they receive training for their part in this study. They were given brief, simple advice on how to sort the 150 essays into nine categories: no criteria or other 'anchors' were provided. Diederich et al used factor analysis to identify groups of readers whose judgements agreed more with each other than with other groups of readers. The readers' comments on the essays were then analyzed in an attempt to interpret the factors which had emerged. The five types of essay readers as described by Diederich et al were:

1. *'Ideas' centred: their comments focused on relevance, clarity, quantity, development and soundness of ideas. Their grades correlated highly with essay length; they were apparently not attracted by unconventional ideas. The readers in this group showed less inter-reader agreement than other groups. Diederich et al suggest this may be because they do not agree on what is a 'relevant' answer.*

2. *'Form' centred: comments focused on analysis and organization, and they appeared very concerned with spelling.*

3. *'Creativity' centred: comments focused on style, interest and sincerity. They were concerned with ideas as were the first group, but seemed to prefer the unconventional to the "merely correct". They gave general comments on mechanics but rarely on specific errors. Four of the seven readers in this group were writers or editors. Diederich et al suggest calling this factor 'Flavour' or 'Originality': it is not exactly 'Style', but one aspect of this.*

4. *'Mechanics': grades in this group were inversely related to the number of errors, with some evidence of adjustment for other types of excellence.*

5. *This group was hard to characterize, although it was not a "ragbag" category because there were high inter-reader reliabilities. Readers gave frequent comments on choice and arrangement of words; Diederich et al suggest this might be an 'Effectiveness' factor.*

There were, in addition, common concerns across the reader types: concern with clarity of expression, coherence and consistency of ideas, and logic (reasoning) showed through in the comments of all the groups, although Diederich et al suggest that they are not really common, but common terms used for different qualities.

Diederich et al pointed out that this was a theoretical study, and had no immediate practical application or normative significance. However, they drew two conclusions from the study:

1. *any reliance on essay grades without computing reliability is almost certain to be unwarranted, since 94% of the essays received 7, 8 or 9 out of the 9 possible grades, and no papers received less than 5 out of the 9 grades: the median inter-reader reliability was .31.*

2. *the readers' task may be simplified by restricting their attention to factors 1, 2 and 3 (i.e., Ideas, Form and Creativity), since factors 4 and 5 could be tested objectively.*

It must, however, be noted that not only were Diederich et al's readers not trained in a common marking method, they were given almost no guidance of any kind (see Section 3.2.). In addition, his readers were from very disparate backgrounds, whereas several studies have emphasised the need for raters to be from as homogeneous a background as possible (Section 3.1.2). Diederich's first conclusion must, then, be open to question.

Although Diederich's second conclusion might be acceptable for the fourth factor ("Mechanics"), it is difficult to accept for the fifth factor

("Effectiveness"). It is difficult to see how a "hard to characterize" category with high inter-reader reliabilities is any more amenable to objective measurement, since presumably Diederich et al are not sure what to attempt to measure.

Freedman (1977) questions the study because it confounds rater qualities and rating technique variables: for example, the papers probably appear in different but not fully randomised sequence in each cluster; and each group of readers is identified only by the most reliable (i e., the three highest and the three lowest) readers, which meant that the comments of 28 of the 53 readers were discounted. An additional weakness is that in preparing the frequency count of the readers' comments, only one judge classified the comments into ideas, style, organization, paragraphing, sentence structure, mechanics and verbal facility. Thus the basis of the description of the groups is itself open to question. Jacobs et al (1981) criticise the Diederich study for a different reason: because it is based on what the readers said about what they responded to in the essays rather than on what they actually did respond to. The same criticism can be levelled at the study by Remondino (1959); the response patterns Remondino identified were very similar to those identified by Diederich et al, with the addition of a factor of readability and appearance

3.1.1.1. Validating reader self-reports

Generally, readers claim to give most weight to the strength or weakness of ideas, content and organisation when grading essays, and less to surface or mechanical features. Freedman (1977) used a research design which did not depend on readers' views of what they were responding to, but which manipulated the essays themelves to try to discover what effects these variations in the essays had on readers. Freedman's research design took into account all kinds of reader features such as educational background, teaching experience, writing experience and "personal problems", and also carefully controlled the environment of the

rating, e.g., training for evaluation, length of time rating, physical environment. She found that teachers' ratings of the essays corresponded well with their statements of the criteria they believed were important in evaluating essays. Harris (1977), in a similar study, found that teachers' ratings generally corresponded quite well with their stated criteria, but that they tended to be more influenced by mechanical features than they claimed to be. She also found that teachers' ratings were more reliable when they were given explicit criteria to use in grading than when they used their own implicit criteria.

The study by Freedman (op cit) showed a clear pattern of readers ascribing most value to essays which were strong in content, with organization having the second most significant effect followed by mechanics, and sentence structure having no significant effect as an independent variable. However, sentence structure contributed significantly together with strong organisation as an influence on scores, as did mechanics. Harris (op cit) also found that readers valued content and organization more than sentence structure, mechanics, usage or "diction" (style), but not as much more in practice as they claimed for themselves. In fact, her raters were quite strongly influenced by factors of mechanics and usage. Freedman's readers were college freshman English teachers, while Harris' readers were high school English teachers, and it may be that the level taught influences the values teachers apply to the ratings of essays.

Problems arise with studies which manipulate the data in order to produce effects in subjects, just as there are problems with studies such as Diederich's which rely on self-reports. Diederich's study is probably the most well-known and influential investigation of writing assessment there has been to date, and is frequently cited, virtually always in support of decisions antagonistic to research in writing assessment through direct writing samples. The dearth of research on raters' processes and bases for judgements, and the unsatisfactory nature of research methodologies

and research questions, led to the ethnographic study reported in Chapter 5, section 2.

3.1.1.2. Handwriting

Chase (1968) looked at the effect of the writer's handwriting on grades assigned by readers, and found that readers were significantly more generous with good handwriting than with poor handwriting. He found, however, that the influence of handwriting was reduced if two papers with poor handwriting did not appear consecutively. Markham (1976) found that elementary school teachers consistently gave good handwriting higher grades than poor handwriting when content was held constant. Soloff (1973) investigated the effect of the handwriting of eleventh grade children on teachers' grades, and obtained a similar result Robinson (1985) found significant interactions between the first language of ESL learners and the responses of raters to their handwriting.

3.1.1.3. Spelling

Chase (op cit) also investigated the effect of spelling on grades, and found no significant effect. On the other hand, Stewart and Leaman (1983) found that the number of spelling errors in an essay was a good predictor of essay grades, a finding supported by Robinson (1985).

3.1.1.4. Length

Grobe (1981) found that on essays written in grades 5, 8 and 11 essay length accounted for between 20% and 30% of the total variance in holistic ratings. Thomas and Donlan (1982) found that the number of words in an essay test answer was the variable most highly correlated with judgements of overall quality regardless of the student grade level. Stewart and Leaman (op cit) found absolute essay length a significant factor in essay grades. Brossell (1983) found that essay length

correlated at p <.001 with score. He suggests intrinsic motivation, previous training, willingness to use the full time limit, and rater bias as reasons for this relationship, but stresses as the main reason the fact that longer essays provide fuller information about the topic.

3.1.1.5. Sequence

Even the sequence of papers can influence the ratings assigned by readers, as Hales and Tokar (1975) discovered. In a study of the responses of 128 pre-service teachers they found that the same answers were scored significantly higher if preceded by five weak papers than if preceded by five strong papers. They find in this support from Helsen's (1950) Adaptation Level (AL) theory, and a number of other studies have tried to account for this by reordering the sequence of papers for different readers (e.g., Jacobs et al, 1981).

3.1.1.6. Writer characteristics

Characteristics of the writers themselves can also influence readers' evaluations: for example, Newcomb (1977) found a significant advantage for female writers over male writers, and for white writers over black writers. Diederich (1974) described a study by Rosner, who manipulated writer characteristics to investigate whether teacher-grades were influenced by student's sex, grade level or stream (honours/regular). He found the only factor for which the graders showed bias was stream: papers believed to be from the 'honours' stream scored one grade higher on average than the same papers when believed to be from the 'regular' stream.

3.1.1.7. Writing characteristics

Thompson (1976) looked at the frequency of various types of "errors" in composition such as lack of unity, lack of clarity and independent judgement errors (for example, flaws in argument) as well as a range of mechanical errors such as incorrect idioms, indenting errors and spelling. He found that the errors most predictive of ratings arrived at by multiple-marking were unsupported statements, independent judgement errors and lack of unity. Dilworth et al (1978) found that papers judged to be superior showed a high level of abstraction/generalization supported by specific information: they demonstrated a superior level of conceptual development together with a good use of syntactic strategies; length of T-units was also greater than in inferior papers. Dilworth et al believe their data reveal a clear relationship between ideation level and syntactic control, a finding which is confirmed by Henning (1982) and Jacobs et al (1981). The findings of Freedman (1977) and Harris (1977) discussed in Section 3.1.1.1 are also relevant in this regard.

3.1.2. Reader effects

Cooper (1977), Jacobs et al (op cit) and others have emphasized the need for readers to be relatively homogeneous for high reader reliabilities to be obtained. Newcomb (op cit) found that rater behaviour varied and could be to some extent predicted according to their background: men were harsher than women; blacks were more lenient than whites; raters from N.E. U.S.A. were harsher that raters from Central U.S.A. But he failed to demonstrate that raters assign higher grades to essays showing their own background characteristics than to other essays. Wesdorp, Bauer and Purves (1982) also showed a significant effect for rater background. Branthwaite et al (1981) found a strong positive correlation between the scores university faculty assigned to written test answers and their own scores on a personality test. Hake and Andrich (1971) used a Rasch analysis model to identify the harshness/leniency of raters and to

predict the scores they would assign essays from this, i.e., they worked from the premise that raters would not agree but that each rater would be self-consistent. Their predictions worked well in most cases, and in the cases where they did not, the misfit could be explained in terms of the interaction between the rater's affective response to an essay and the essay topic/treatment. (A rater who had recently miscarried scored a pro-abortion essay lower than predicted, and so on.)

3.1.3. Training effects

Stalnaker (1934) showed the importance of training for improving essay rating reliability. In his study, inter-reader reliability varied from .30 to .75 before training, whereas after training it improved to between .73 and .98 with an average of .88. Pilliner (1968), Coffman (1971), Diederich (1974), Jones (1975), Cooper (1977) and Jacobs et al (1981) all stress the importance of rater training, although they do not provide empirical data. Freedman (1981) not only found significant effects for training (second largest effect after the essay itself), but she found that the trainer who conducted the training session had a significant effect: trainers who discussed the topic in detail with their raters prior to rating led the raters to award significantly higher scores. Daly and Dickson-Markman (1982) found that scores assigned by raters showed a strong positive influence but not a negative influence for the essays used in the training session prior to the reading. The elaborate training procedures used by ETS in rating essays (described, for example, in Carlson et al, 1985) depend to a great extent on standardization of raters, and raters work physically together: their scores are carefully monitored and they are kept 'on scale' by continual restandardization. Carlson et al (op cit) report that "the careful training procedures in this study were sufficient to overcome any differences in rating strategies between ... types of readers...". However, Newcomb (1977) found little evidence in his survey of the literature to provide empirical support for the efficacy of training. He investigated training effects, and found that after training

his raters achieved only an average inter-rater reliability (i.e., single-rater reliability) of .72. He found that after training raters still applied their own criteria when rating (the following day).

There do not appear to be any studies of the effect of training in contexts such as that used in Diederich et al's study, or that which is the focus of the study reported in Chapter 6, where a wide range of raters each work in isolation, and the training they receive is through print materials, but it seems unlikely that training would be as effective, or inter-rater reliability as high, as when raters work physically closely together.

3.2. Procedural variables

Cooper (1977) divides scoring procedures for direct essay tests into two broad categories, 'holistic' and 'frequency count'. He describes 'holistic' methods of evaluation as "any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing" (p.4). For Cooper, 'frequency count' methods include error counts, T-unit counts and other methods in which features of the writing are enumerated. As Cooper admits, there is some difficulty in distinguishing the most atomistic of the 'holistic' methods from the most integrative of the 'frequency count' methods. Mullis (1984) classifies methods of scoring direct writing assessments into three: holistic scoring, primary trait scoring, and analytic scoring. She appears to discount frequency count methods altogether. In the Section which follows scoring procedures are classified as holistic, analytic or frequency count.

3.2.1. Holistic scoring

Holistic scoring methods involve reading the essay for an overall impression of the quality of the writing and assigning it a score based on this overall or 'global' quality. Holistic scoring is based on the

view that the quality of any piece of writing is greater than any of its directly observable parts, but that this undefined quality is something that skilled readers can recognize. Typically readers are asked to read rapidly, forming a global judgement and not focussing on any specific features such as organization, mechanics or ideas.

3.2.1.1. Impression marking

Impression marking is the simplest of the holistic scoring methods for direct tests of writing. In impression marking the reader makes a judgement based on an impression of the composition as a whole, without attempting any analysis or break-down of the writing into features. It is essentially a norm-referenced procedure, since most commonly the rater decides where each paper fits within the range of papers produced for a given rating occasion (Cooper, 1977). The essence of impression marking is that raters should read each essay quickly, forming an impression and assigning a score, and not re-reading or reconsidering their score. This is the method which was used in scoring traditional examinations, until a variety of other scoring methods were introduced in attempts to reduce the unreliability shown to exist in impression marking.

In the last twenty years impression marking has been developed and refined, and coupled with some of the other techniques discussed below, creating a range of holistic evaluation procedures, the best known and most researched of which is that originally developed at ETS by Godshalk et al (1966) and refined over the period since then. Coffman (1971) justifies impression marking by saying:

> *To the extent that a unique communication has been created, the elements are related to the whole in a fashion that makes high interrelationships of the parts inevitable. The evaluation of the part cannot be made apart from its relationship to the whole. (p.293)*

Lloyd-Jones (1977) takes this a little further:

> *One need not assume that the whole is greater than the sum of its parts - although I do - for it may simply be that the categorizable parts are too numerous and too complexly related to permit a valid report. (p.36)*

However, Cooper (1974) argues that even when readers claim to be wholly impressionistic they very often use a set of implicit criteria which they can make explicit if required.

3.2.1.2. Essay scales

An essay scale is a series of candidate essays, at least one for every point on the assessment scale, the purpose of which is to exemplify the standard or criterion for each point on the scale. Willing (1918) and Van Wagenen (1920) experimented with essay scales in the early years of the century, but until recently essay scales were not much used as criterion measures in large scale writing assessment. LATE (London Association of Teachers of English) developed an essay scale (Martin et al, 1965) which consists of multiple examples at each point, together with a discussion of the strong and weak points which raters had found in them.

The holistic evaluation methods developed from the work of Godshalk et al (1966) for ETS, and used for the evaluation of the Advanced Placement Test in English (Smith, 1975), the National Assessment of Educational Progress (NAEP, 1980) and the English Composition Test (Conlan, 1978) include the use of essay scales consisting of 'rangefinder' or 'benchmark' papers, but these are developed anew for each test topic and rating session. The scoring procedure for the TOEFL Writing Test (Stansfield & Webster, 1986) uses the 'rangefinder' system. The California State University system has used an essay scale since 1973 as part of its essay scoring procedure (White, 1985), and essay scales are increasingly

used in American school systems, but in general these, like the ETS essay scales, are developed for each rating occasion.

3.2.1.3. Focussed holistic scoring

Focussed holistic scoring is a scoring procedure which is holistic in orientation but which is focussed on a number of objectives seen as central to the purpose and context of the assessment (Sachse, 1984). Focussed holistic scoring is usually based on a list of features expected to be present in the writing being scored, or a list of things the writer must demonstrate that she can do in writing. Readers score essays with reference to the feature list but do not look in depth at each feature as it appears in the essay: judgements are reported as one global score. Although not confined to this, focussed holistic scoring is particularly suited to criterion referenced evaluation of writing.

3.2.1.4. Rating scales

A rating scale closely resembles a feature list but the features are arranged into a prose description, usually a short paragraph, and the features are described in terms of the performance they can be expected to be represented by at the various levels on the scale. The TOEFL Writing Test (Stansfield and Webster, 1986) uses a rating scale with six points as part of the holistic scoring procedure.

3.2.2. Analytic scoring

Analytic scoring is distinguished from holistic scoring in that prominent 'characteristics' or 'features' of the writing in the context are not only identified but also scored separately. Rather than a global score, each essay receives several scores. Analytic scoring is based on the view that, although the overall effect of an essay may be different from the sum of its parts, the characteristics of an essay can be described in

meaningful ways, providing an analysis of its strengths and weaknesses. Analytic scoring provides a record of why the paper received the score it did. Analytic procedures vary considerably, from those which are very similar to feature lists or holistic rating scales, to those which approach frequency counts. Unlike holistic procedures, where it is relatively difficult to focus the basis upon which the reader is to make a judgement to reflect the context of the writing and the purpose of the asssessment, analytic procedures can be developed to suit different contexts and purposes.

3.2.2.1. Analytic categories

The analytic scale developed by Diederich (1974) is typical of the less specific scales: it contains the main categories 'General Merit' and 'Mechanics'; the first is broken into Ideas, Organization, Wording and Flavor; the second is broken into Usage, Punctuation, Spelling and Handwriting. Each of these is scored on a 2-4-6-8-10 system, but none of them is defined or described. Pike (1973) developed a scale which consisted of four sub-scales: content quantity (number of ideas and concepts and degree of elaboration); content quality (adequacy/interest of story line; internal consistency; flow); form quantity (range of vocabulary and grammatical structures); form quality (appropriacy and effectiveness of vocabulary and grammar). Currently analytic categories are in some disfavour and analytic scales are preferred.

3.2.2.2. Analytic scales

Analytic scales are distinguished from analytic categories by the fact that they contain a description of each feature or characteristic at each point along the scale. These descriptions may simply be adjectives or adjectival phrases. Chaplen (1969) used an 'Essay Marking Scheme', consisting of six points from 'Excellent' to 'Hopeless', each with a brief paragraph description of typical performance. Palmer and Kimball (mimeo,

undated) developed a 'Criterion-Based Composition Grading System' consisting of nine characteristics each described on a very simple scale (e.g., Propositional Content: Statements are comprehensible on a single fast read only: 2-Consistently comprehensible; 1-Most comprehensible, though some are not; 0-Generally incomprehensible). The 'Composition Profile' of Jacobs et al (1981) uses a fully described analytic scale. Developed for use in scoring college-level ESL writing, the Profile has been carefully worked out and extensively validated. It comprises five components (content; organization; vocabulary; language use; mechanics) each of which exists at four levels and a brief indicator of the characteristics of writing at each level in each component is provided. Mullen (1980) used an analytic scale consisting of five sub-scales: control over English structure; organization of material; appropriateness of vocabulary; quantity of writing; and overall writing proficiency. Henning (1982) developed an analytic scale consisting of five mechanical criteria and five content-oriented criteria. He found that scores which combined the mechanical and content ratings were more reliable than scores based on only one set of criteria. The analytic scale used by Weir (1983) consists of seven 'attributes', each of which is described on a four-point scale. Braddock et al (1966) point out that an analytic scoring method is of little use if the criteria it contains are not well-defined: for example, "quality of ideas" is not a well-defined criterion. Because it is so much more difficult to define criteria such as intellectual quality, rhetorical effectivenes and fluency than mechanical criteria, these criteria have tended to be under-represented and vaguely described in analytic scales, which leads to their under-weighting in a final score and to a lack of validity for the overall score.

3.2.2.3. Dichotomous scales

A dichotomous scale, as its name implies, is a series of statements which the rater can answer 'yes' or 'no'. Cohen (1973) describes an experiment in which twenty one English instructors worked together to develop a

protocol consisting of nineteen items which they agreed reflected the quality of an essay. The amount of agreement between the instructors as to whether an essay demonstrated each of these qualities or not (i.e., a dichotomous scale) varied from .50 (chance) to 1.00, while an analysis of the reader reliability on each quality showed much more agreement over some qualities (creativity; number of modifier errors; appropriacy of paragraph development) than others (thesis development; clarity; organisation): there was a tendency for there to be more agreement over mechanical qualities than content organization. Cohen states that each of the sub-scales on his scale correlated well with the total (with or without themselves is not stated) but poorly with each other, and suggests that "each scale is measuring an independent variable" (p.365). The instrument developed by Hake (1973) and Andrich (1973) includes a complex dichotomous scale.

3.2.3. Holistic vs. analytic procedures

What all analytic scales have in common is the attempt to focus more precisely on the qualities of the writing which are important for the purpose of the assessment. This attempt at precision is in contrast to the imprecision which is counted as a virtue by the proponents of holistic scoring. Brown (1981) says.

> *No matter how reliable holistic scoring is as a way of rank-ordering papers, it is inadequate as a measuring tool in itself, because it is relativistic and is not tied to any absolute definition of quality.*

He goes on to recommend the use of a rubric or protocol (i.e., an analytic scale rather than just analytic categories, in the terms used here) to ensure the absolute standard he believes to be necessary. Hirsch and Harrington (1981) are in agreement:

> *Since we differ among ourselves in valuing writing, we cannot expect to achieve uniformity of judgement when we are rating the quality of writing....<u>unless</u> we agree to*

> *adopt special criteria on the basis of their overwhelming rational appeal....valid assessment of writing will depend finally, on our being able to distinguish those significant qualities of writing about which people can agree, from those qualities about people cannot possibly agree. (p.194)*

From the above it is clear that a fully developed analytic scale is an attempt to establish a stable criterion against which each writing sample can be measured. Not all researchers are agreed about the virtues of analytic scales, however. White (1985) briefly discusses analytic scales and is scathing in attack of them:

> *Analytic scoring is uneconomical, unreliable, pedagogically uncertain or destructive, and theoretically bankrupt. (p. 124)*

The only analytic instrument cited by White to explain his extreme reaction is Cooper and Odell (1977), described in Section 3.2 2.1. The more recently developed analytic scales (e.g., Jacobs et al, 1981, Weir, 1983), as we saw in section 3.2.2.2., can withstand White's attack.

A further argument made against analytic scoring (Cooper, 1985) is that it may create a 'halo effect' i.e., raters may in fact be rating for one 'general impression' factor, but doing so repeatedly. If the components in an analytic scale were indeed measuring the same thing, a scale would be unnecessary, since it would contribute nothing extra to the total score information. This view is understandable in cases such as Pike's (1973), who found that his two Content scales showed no distinction and therefore combined them: the same occurred with the two Form scales. The finding that readers do not distinguish quantity of and quality of the features described in the scales is perhaps not surprising, and suggests that the components in an analytic scheme need to be more carefully established than these were. More recently, however, the procedures developed have been carefully validated to ensure that they are free of halo effect.

Jacobs et al (op cit) found that their components correlated at between .57 (mechanics with organization) and .88 (vocabulary with language use). The .57 correlation suggests that these two components are measures of somewhat different things, an intuitively satisfying conclusion, as is the much closer correlation of .81 between content and vocabulary. Weir (1983), in a rigorous study, found no statistical justification for the combining of any of his seven attributes. Kroll (1982) showed that there were consistently weak correlations between scores assigned for discourse level features in a set of 100 test essays and scores assigned for syntactic level features. It may be that some candidates will show parallel command of all the facets of writing skill, while others will show uneven development. Clearly, the finely tuned assessment of an analytic scale is needed in order to find this out.

Woods Chapman, Fyans and Kerins (1984) recommend the use of analytic scoring for other reasons: they consider that an analytic scoring method is more reliable:

> *While each particular writing item (focus, support, organization, mechanics) has its own unreliabilities and invalidities, taken together, they are quite powerful in describing the student's ability. ... The high loadings on each scale (.70 to .90) were all on one factor, thereby supporting the aggregation into one writing ability score. (p. 25)*

Purves (1984) found that the addition of an overall-impression category, 'Personal Response of the Reader' to his analytic scoring procedure was particularly acceptable to readers and "appeared to lessen what is known as 'the halo effect' " as well as appearing to enhance the agreement of the readers on the other categories (p. 437).

3.2.4. Primary trait scoring

Primary trait scoring falls between holistic and analytic scoring. In this procedure, criteria are clearly defined and levels are specified, as

in analytic scoring, but only one judgement is made, as in holistic scoring. According to Odell (1981), primary trait scoring "rests on the assumption that different tasks, even different expository tasks, may have to be judged by different criteria" (p.124). Primary trait scoring starts with a clear description of objectives. The writing task is stated as a form of communication involving purpose, audience, and subject, and the unique features of writing to be elicited by the specific writing task are described in scoring scales.

Cooper (1985) describes primary trait scoring as neither norm-referenced nor set to minimal competency standards (criterion-referenced), but as starting from an idea of the best student model at the level being assessed. Primary trait scoring tends to focus the attention of the rater on discourse level features, such as the number and placement of propositions and their supports: Gere (1980) has suggested that primary trait scoring should be supplemented with frequency counts of other important features. Lloyd-Jones (1977), who introduced primary trait scoring, makes it clear that it must be based on a carefully worked out theory of discourse, since the assessment must be related to the purpose of the writing. It is, in fact, a context-specific assessment instrument. Henning (1983) suggests that primary trait scoring may be difficult for teachers to work with, because they are accustomed to working with a single procedure whatever the mode or purpose of the writing, and also because few models for primary trait scoring have as yet been worked out. These are not arguments against its use in large-scale assessment programmes, however, and primary trait scoring has been implemented by, for example, the NAEP (National Assessment of Educational Progress) with some success (Mullis 1980).

3.2.5. Frequency counts

There is a wide range of different features of written text which can be counted in order to arrive at 'objective' measures of writing quality. For example, Jurgens and Griffin (1970) looked at seven objective measures of language production: number of words; number of T-units; number of subordinate clauses; number of clauses generally, words per clause; words per T-unit, and clauses per T-unit, and found that quality was distinguished most clearly by the number of words (i.e., length), number of T-units, and number of clauses of all types. Their criterion was the average of two holistic evaluations by graduate students in English, but correlations between criterion and objective measures were not reported. Page (1967) developed 30 predictor variables by which a computer could attempt to predict reader rating. His best predictors were:

1. standard deviation of word length through essay	*.53*
2. word length	*.52*
3. number of common words (according to Dale list)	*-.48*
4. essay length	*.32*
5. number of spelling errors	*-.21*

None of these predictors seem inherently meaningful as a measure of writing proficiency, nor is any of them particularly reliable (the best of them is at the lower boundary of inter-reader reliabilities generally reported). Perkins and Leahy (1979) looked at number of error-free T-units, error-free T-unit length, and the ratio of clauses per error-free T-unit as predictors of holistic evaluations of the writing of native and non-native undergraduates. They concluded that none of these measures could consistently differentiate between the two groups whereas holistic evaluations "seem to go far beyond what can be measured through the use of these objective measures" (p.310).

Hake (1973) developed a different approach to essay marking which nevertheless falls into this category. Working from a Chomskyan theory

of language learning, and seeking to develop a method to which a colleague (Andrich, 1973) could apply latent trait theory (Rasch analysis) in order to objectively evaluate various essay components, she hypothesized that "the formulated whole generates its parts" (p.35) and that "to attempt to judge the whole, we must first determine if the whole exists and then determine if its parts are functioning to communicate" (p.37). She posits four dimensions for any essay, the first of which is its 'deep structure' and the other three are different aspects of 'surface structure'. The scoring method assumes that the whole is the sum of its parts: in the scoring procedure the scores on the three surface dimensions summed equal the other, i.e , deep structure dimension. The procedure involves counting each occurrence of an error (called by Hake and Andrich a 'flaw')and categorising it, the aim being to achieve 'grader-free' measurement.

However, the study by Andrich (op cit) shows that grader-free measurement did not result. On the Rasch analysis 'misfitting' essays were not the same for each grader, and there was interaction between essay and grader and between grader and dimensions. There is a problem in the use of Rasch or any other latent trait method in that the assumption of stochastic independence is not true for flaws in essays, i.e., a writer who has not mastered a linguistic form will repeat the same error - the errors are clearly dependent. Similarly, a rater's flaw count will show consistent relationships among flaw observations. It might be possible to describe each grader in terms of harshness, leniency, etc., but it does not seem possible to make any useful predictions on this basis. Further, the grader-dimension interaction led Andrich to suggest that their hypothesis that the essay is one trait (i.e., "an integration of specifiable but collaborating dimensions", Andrich, op cit: p.5) is not upheld: some writers appear to be worse on some dimensions than on others.

Flahive and Snow (1980) investigated correlations between objective measures and holistic evaluations of ESL writing at several proficiency levels, and found only one correlation above .70: while considering that they had demonstrated that objective measures were relatively useful in determining levels of overall ESl proficiency and of writing ability, they concluded that "there is far more to writing than length-of-T-unit or clause/T-unit ratios" (p. 176). After surveying a range of studies of objective measures of writing ability, Perkins (1985) concludes that:

> *the use of objective measures is impractical, tedious, and time consuming for classroom use; while objective measures may be of interest to researchers, they, seemingly, are of little value in assessing the underlying constructs of writing because the intent to communicate is neither assessed nor measured by them. (p. 662)*

3.2.5. Effects of scoring methods

All of the scoring methods surveyed above are attempting to get at the same thing: proficiency in writing. Thus they should all result in comparable scores. However, there are few empirical studies of the influence of the choice of scoring method on actual scores assigned to writers, or on reader reliability levels. Winters (1979) used four different scoring methods to assess Freshman essays at UCLA: the Diederich expository scale, the Center for Study of Evaluation (UCLA) procedure, impression scoring, and T-unit analysis. She found that all the methods were adequately reliable but that they produced different patterns of results. Some methods showed variation for topic differences while others did not: impression scoring did and T-unit analysis did not. Winters considers that all scoring systems have a limit of generalizability of which their users should be aware, and she suggests scoring by several methods and combining the results, to control for method effect. She believes that "at the heart of all scoring systems is a conceptualization of the construct 'writing skill', which may differ

from method to method". The suggestion of using several methods is also made by Mullis (1984).

Discussion of the relative merits of scoring methods centres on the issue of what the construct is which we are trying to measure: impression scoring is based on the belief that the individual response of every reader to a piece of writing is valid, but impression scoring has been shown again and again to be very unreliable. It was to counter this unreliability that multiple marking was developed.

Wiseman (1949) argued that score replication, the combination of these individual judgements of a number of raters, enabled the assessment to get at the underlying writing proficiency of the writer, and this is a view which has been accepted and which continues to inform the conduct of most large-scale writing assessment, such as the ETS and California State University assessments referred to above. Cox (1968) criticized multiple marking on the grounds that while statistical reliability is improved, the improvement does not necessarily represent greater agreement on the actual writing, that is, it does not lead to increased score validity. Pilliner (1969) showed that if there is some initial agreement among raters about the merits of an essay, score aggregating will increase both score reliability and score validity; if there is no initial agreement among raters, score aggregation will increase score reliability without increasing score validity. Where raters cannot agree at a basic level, Pilliner suggests that more analytic methods may be more appropriate.

Decisions about scoring methods must take into account all factors which are considered to be meaningful in the context of the specific assessment. The limited evidence available suggests that a choice of scoring method does not only have effects on the reliability attributable to the assessment, but also to its validity.

Holistic scoring is typically claimed to have great validity, since the essay rater is responding as a reader in the post-structuralist sense of a creator of a text from the text before her (Fish, 1980), and White (1986) describes a holistic reading as the creation of an interpretive community in Fish's (op cit) terms. But this view assumes that validity is defined by the 'authenticity' of the reader's response and nothing else: the problems with this view are not only the statistical ones we saw in the preceding paragraph. With such a view, it becomes impossible to define any of the usual parameters of the assessment and thus to know when any individual reader is responding outside the 'community', or why. When a strong interpretive community is formed, scores will be very reliable, but such a community will find it difficult to accept and initiate new members. Further, the writing of the writer, the writer's interpretation of and response to the task, has no value of its own. In such a view, readers are empowered while writers are dispossessed. Neither does the task have any intrinsic significance, but only the significance the reader (or, in multiple holistic reading sessions, the community of readers) is prepared to ascribe to it.

In analytic scoring and primary trait scoring these problems are lessened and their effects can be recognized and measured. With these methods, the validity problems centre on the critical need to ensure that the purpose and values in the writing context are fully clear to the writers and that the criteria for assessment and the guidance for the recognition and evaluation of the characteristics of the writer's writing are appropriate and accurate. When these requirements are satisfied, analytic scoring and primary trait scoring lead to the empowerment of writers who create successful texts for the defined context, and also to the empowerment of readers who understand what it is they are asked to do and why.

3.3. Writer Variables

The term 'writer variables' is used in this Section to refer to those variables which cause the individual writer's performance to vary from task to task, and from occasion to occasion on the same task, but which are not due to writer unreliability, i.e., they represent true rather than error variance.

The true variance implied by the term 'writer variables' is variance intrinsic to the writer or to the interaction between the writer and the task. It has only recently been recognised that all writing, including the writing of expository-type essay answers to test questions, is creative and personal as well as communicative. This perception has increased our awareness of the many variables, apart from the reader variables we investigated in the preceding section,which can affect an individual's performance: as we shall see, because of the recentness of attention to these aspects of writing assessment, none of them is well investigated or understood.

There is not yet a research methodology for the study of writer variables or their impact, nor even a developed categorisation of writer variables as distinct from the task variables which are the subject of section 3.4. We are at present unsure of the amount of true variance in a writer's essay test score which is due to theoretically predictable characteristics of the writer, or to the interaction between such characteristics and specifiable characteristics of the test task. Brossell (1986) says:

> *All writers are influenced in writing assessments by innumerable factors related to background and personality. Elements of culture, gender, ethnicity, language, psychology and experience all bear upon the way different people respond to a writing task. Unfortunately, the current level of knowledge about such influences does not allow us to understand the precise ways in which human factors affect writers and their performance on writing assessments. (p. 175)*

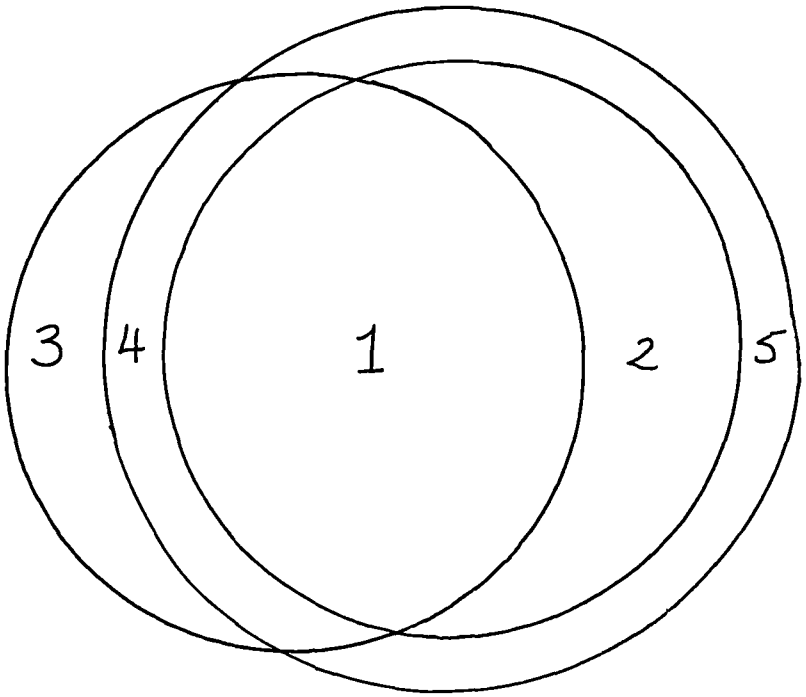
3.3.1. Topic/task interpretation

Each writer brings the whole of himself or herself to the essay test. Each writer is a complex of experience, knowledge, ideas, emotions and opinions, and all of these things come with him or her to the essay test. In interpreting and responding to the topic of an essay test, each writer must create a 'fit' between her world and the world of the essay test topic and she must interpret the task in terms which make sense to her before she can respond to it. When the topic/task is a very wide one, it is, as Labov (1969) said, "absurd to believe that an identical "stimulus" is obtained by asking everyone the "same question". (p. 108) When the task is a very narrow one, writers whose personal histories are the closest 'fit' to the expectations of the task will find it easiest to interpret. Brossell (1983) found that topics with a moderate level of rhetorical specification (specification of purpose. audience, voice, and content) yielded higher mean scores than essays with either a high level of specification or a low level of specification. Each writer needs both guidance as to what is important about this writing task and what qualities will be valued, and some room in which to maneouvre in taking the task and topic and creating an original, personal, response. Pollitt and Hutchinson (n.d.) refer to this as 'outcome space'. They show that there are five possible 'outcome spaces' for any examination question (Figure 2.3.1.).

The writer will always interact with the topic in ways which we do not, at least at present, and for the foreseeable future, understand. Weaver (1973) found that writers need to transform a 'teacher-initiated' topic into a 'self-initiated' topic: that is, each writer must take the task and somehow make it her own. If the testee cannot satisfactorily follow the steps of attending to, understanding, and valuing the topic, she will replace it with another, or with a related, topic, but will not respond to the topic intended. In order to transform the topic, i.e., value it and

begin to respond to it, the task must seem realistic, appropriate and feasible to the writer (Rosen, 1967). Once the topic is accepted, the writer must either do a mental scan for input or write for discovery (Weaver, op cit). Brand and Powell (1985) found that the attitudes of good and poor writers to being asked to write differ: good writers feel fairly positive from the outset, while poor writers feel negative. However, both good and poor writers tend to feel more positive about the fact of writing as they respond to a topic.

Figure 2.3.1.: Five Outcome Spaces



	Acceptable	Observed	Anticipated
1	+	+	+
2	+	-	+
3	-	+	+
4	+	+	-
5	+	-	-

3.3.2. Writer as thinker

A number of composition researchers, for example Elbow (1973) and Holland (1976), have applied Piagetian stages of intellectual development (as promulgated by Flavell, 1963) to a description of the levels of thinking employed by student writers in their composed products, and Holland (op cit) has suggested that writing tasks should be designed to be accessible simultaneously to all of Piaget's four levels of thinking, in order to "create the possibility of discovering at what stage a writer is able to perform with this task" (p. 20). But Piaget's theory was initially developed from a series of simple experiments, and his conclusions generalised to other areas without empirical testing. Peel (1971) explored systematically and empirically the nature of adolescent judgement in a range of content areas: he found that most pupils' responses to a problem-solving task could be categorised into one of four types: mentioning (tautological, partial or inconsistent responses); describing (mainly correct listing of aspects of the topic); explaining (use additional related ideas to interpret textual meaning); combining (combine more than one piece of evidence with outside ideas to interpret textual meaning). Sutherland (1982) similarly analysed responses to biological questions, and developed an eight-category system:

0 completely naive
1 pre-describer (grossly inaccurate or irrelevant)
2 elementary describer (very simple/ limited description)
3 describer (some grasp of underlying concepts)
4 extended describer (use of abstract concepts but not as explanations)
5 transitional (occasional flashes of explanation)
6 full explanations in conceptual rather than perceptual terms
7 extended explainer (explanation in terms of theory and deduction from evidence)

Sutherland found pupils to be at more or less the same stage for every answer: variation for any pupil was rarely more than one either side of her or his own mean. Peel similarly described his categories in terms of a series of developmental stages, but he recognised that the distribution of responses was affected by the nature of the text used to elicit data, by its subject matter, and by the form of the questions used. However, Marton (1981) argues that what should be classified is the application of a particular thinking pattern in a particular task, not the description of the developmental level of an individual. He suggests that performance levels should be expected to vary across tasks. Schroder et al (1967) present an information-processing model which suggests four gradations of information-processing in persons, based on the number of dimensional attributes of information perceived, and the number organized and processed:

> 1 *low integration (relatively fixed or hierarchical organization regardless of the number of dimensions: characterised by black/white thinking)*
> 2 *moderately low integration (emergence of alternate combinations of dimensional scale values: characterised by primitive internal causation, instability and noncommitment)*
> 3 *moderately high integration (alternate perspectives; more complex rules for comparing and relating: characterised by empiricism, awareness of choices)*
> 4 *high integration (structure for generating complex relationships: characterised by self-reflection)*

Schroder et al point out that these are not discrete steps but occur along a continuum: the model is used to score answers to essay questions on a psychology examination, on a seven-point scale, where points 1, 3, 5, and 7 correspond to the four stages in the model (above), and 2, 4 and 6 are degrees between the points. Schroder et al say:

> *... essay questions that produce the most construct-relevant responses have the following components: (a) they present the student with uncertainty and conflict; (b) they express a point of view and ask the student to consider his agreement or disagreement with it; (c) they present two discrepant points of view; and (d) they*

> *present a number of ideas, that, presumably, have not been related for the student beforehand, and about which he is asked to consider relationships. (p. 200-201)*

Schroder et al implied that respondents who show high integration can operate at the lower levels on different tasks, but that respondents who show low integration on one task cannot operate at higher levels. Their work was adapted and developed by Biggs and Collis (1982), who, in contrast to Schroder et al, describe performances rather than persons, and explicitly recognise that there will be different levels of performance on different tasks. This viewpoint is reiterated by Pollitt and Hutchinson (n.d.), and brings us once again to the apparent impossibility of separating writer variables from task variables.

3.3.3. Writer as community member

Wilkinson (1983) criticises models of writing development which are restricted to descriptions of linguistic skills or cognitive abilities. The model developed by Wilkinson and his team includes 'affective' and 'moral' measures as well as the more usual stylistic and cognitive measures. In this model, audience awareness is part of the affective measure, and the moral measure is seen as the internalizing of the morality of the culture. Concerned as it is with the language development of young schoolchildren, Wilkinson's model does not directly treat social development, although through the descriptions of the stylistic, affective and moral characteristics a picture of the developing writer's social development may be obtained.

Many of the ways in which researchers into the composing process and judges of composed products are currently looking at and responding to writing have a social perspective. Faigley (1986) identifies a "social view" (p. 528) as the most recent of what he calls in his title the "competing theories of (writing) process". The extreme of this view is represented by social constructionist thought, as most frequently

encountered in composition theory in references to Kuhn (1970) and Geertz (1973; 1983). Kuhn believes that all scientific knowledge is a social construct, whereas Geertz believes that all knowledge is a social construct.

A social view, if not a social constructionist view, of writing underlies the teaching of writing for academic and specific academic purposes. Herrington (1986) identifies three social communities within which the student of an academic discipline functions as an apprentice writer: the educational community; the disciplinary community; the classroom community. Clearly, any apprentice writer as a member of a classroom community is also a member of a wider home and social community, which to a greater or lesser extent supports the writer as she is inducted into the more specific communities of institutional education, depending on the extent to which all these communities are in harmony. When the writer perceives herself as a member of a community which is in disharmony with the wider community, the writer may experience stress, seeing the values of one community as in conflict with those of the other.

Within one 'culture' (in political or geographic terms) we typically find different communities or sub-cultures which share many ways of seeing, doing, and thinking which they do share with the larger culture. Luria (1976) reports on tests conducted in the USSR in which traditional, illiterate, peasants were compared with more educated, literate, members of local communities on the categorizing and sorting of everyday objects and geometrical figures: the illiterates were shown to be more situation-bound, more concrete, and less likely to use abstract language or to make generalizations than their literate neighbours. Gere (1981) tells us that cultures may share different views of the value of literacy: she speaks of Black American culture, which places a high value on oracy skills and where members may prefer to remain illiterate and retain the "inner link of community" (p.120). She also suggests that Native American

communities may resist literacy programmes because reading and writing "threaten the kinship they value" (loc cit).

The writer's sense of herself as a community member is part of what the writer brings to a writing test, as it is a part of what she brings to her perception of and value attributed to writing as an art or skill. Giannisi (1976), in a bibliographical article, reviews a number of studies in Standard English as a Second Dialect and English as a Second Language relevant to the question of dialects and the teaching of composition, and concludes that "greater understanding of varieties differentiation is crucial to the teacher of composition in a pluralistic society" (p. 283). Toelken (1975) quotes a Navajo Indian student writing in a Navajo newspaper:

> *The big education conferences and workshops will beat around the bush and make excuses about why a high rate of dropout. ...One thing nobody even mention is Indian students drop out because they are Indian, thinking Indian loud and clear, but you can't hear him. He will get fed up with trying to learn what he does not believe. Eventually, he quits or flunks out at the end of semester.*

Toelken explains that the writer cites a worksheet used on Navajo reservations to teach the meanings of various English lexical items by showing the lexical item in cartoons, which had been 'translated' into Navajo culture. The learner is supposed to mark one of two pictures right or wrong, to match the statement "The father works": one picture shows the man chopping wood; the other shows him leaning on a building smoking. The student says:

> *...nothing is said about the fact that "working" is an entirely white idea. Only the whites divide their time so carefully between "work" and "play". You have got to work or play. "Just hanging around" is "wasting" time and doesn't count.*
>
> *Of course, Navajos have entirely different ideas about time, and about work and play. This is the real Cultural Difference, and not what sort of skirt the mother happens*

> *to wear, not whether the father wears his hair short or tied up behind.*
>
> *In school, the kid who happens to know how white people feel about work will be able to mark this worksheet in no time flat, will get a good grade, will be promoted and all the rest of it. The Navajo kid who is confused will fail, "fall behind" - and will end up "no good". (p.279)*

Similarly, writing tests typically place a high value on the use of 'standard written English', and in the United States such tests have been consistently found to result in significantly higher test scores for whites than for blacks (White and Thomas, 1981).

3.4 Task Variables

From the discussion of reader variables and writer variables, it can be seen that it is very difficult to find ways to shape writing tests to give every testee an equal opportunity to give her best performance. In this Section we shall look at task variables, that is, those aspects of the writing test task which can be, at least theoretically, deliberately manipulated, and the manipulation of which can cause differing performance by testees.

3.4.1. Content

Hartog (1936), Hartog et al (1941), Braddock et al (1963), Britton et al (1966), Poetker (1977), Hirsch and Harrington (1980), and Applebee (1983) have all stated that content quantity and quality in an essay answer are directly attributable to the topic of the question. Hartog (1936) had early suggested that topic familiarity or unfamiliarity would affect a writer's performance, and had felt that the solution would be to limit the topic by defining a purpose and audience for the writing and by not permitting a choice of topic. This was a major recommendation of the report by Hartog et al (1941). Weir (1983) and Bridgeman and Carlson (1983) have shown that content criteria are the foremost criteria for

determining academic faculty grades on expository test essays. Hirsch and Harrington (1980) found that the communicative effectiveness of compositions was based on the writer's familiarity with the essay topic. Freedman and Calfee (1983) found that essay scores were significantly influenced by the subject matter of the topic. In contrast to the other findings reported here, Brossell and Hoetker Ash (1894) studied 21 different essay topics and found their content to be of slight consequence to scores.

There have been attempts to circumvent the problem of the topic variable by asking students to write about anything they wish, or by providing a content-neutral topic such as 'Red' or 'Yesterday'. However, Scardamalia et al (1982) remind us that although writing is unlike conversation in that each 'turn' (e.g., a sentence) does not get a response, an input in the form of an external signal, it is nevertheless a speech act. They report a study which showed that even feeding children contentless external signals while they were writing stimulated them to write 70% more functional text. This provides support for the common practice of providing a certain amount of stimulus or input material to flesh out the essay topic. Applebee (1983) goes further when he suggests that topics should be very familiar to the writer:

> *When the topic being written about raises questions that have not been fully explored in the past, the written language may become a tool for ordering and clarifying relevant knowledge and experience. In this case, the writing task becomes a heuristic one, a process of discovery and reformulation which will in all likelihood be halting and somewhat uncertain. (p.367)*

Certainly Applebee's proposition accords with a view of writing as both a channel of communication and a vehicle for discovery of knowledge, and it also accords with the research suggesting that some modes of discourse are more difficult than others. A study of high school students' self-evaluation of their knowledge of topics in chemistry (Johnstone et al, 1971) showed that when students considered themselves not to have

mastered a topic in the syllabus, their test scores on that topic were correspondingly lower than their test scores on topics they felt themselves to have mastered. Pollitt et al (1985), who investigated high school students' responses to questions in geography, mathematics, chemistry, French and English, found that subject difficulty was an important influence on student performance, and located part of that difficulty in the degree of familiarity of the students with the concepts (i.e., with vocabulary items to express concepts) used in the questions.

3.4.2. Purpose

Witte et al (forthcoming) found that college freshmen benefited from a clear specification of purpose for the writing test: as purpose was more exactly specified, scores increased.

3.4.3. Audience

The audience variable has also been shown to affect a writer's performance markedly. Smith and Swan (1977), Rubin and Piche (1979) and Crowhurst and Piche (1979) all found significant effects on the quality of writing as a result of the specification of different types of audience. To the extent of their writing ability, writers adapt the way they write to their perception of the expectations and capabilities of their audience.

Smith and Swan (1977) took Hartog's (1936; 1941) suggestion of specifying the audience for an essay task, and conducted a study which showed that differing target audiences affected college students' writing, but not the writing of sixth grade children, suggesting that audience adaptation is a late-blooming skill. Rubin and Piche (1979) investigated the effect of differing target audiences on the writing of fourth, eighth and twelfth grade children. They asked the children to write for audiences of high,

medium and low intimacy, and found significant differences both semantically and syntactically at the higher levels.

Crowhurst and Piche (1979) looked at the effect of differing target audiences on the syntactic complexity of the writing of sixth and tenth graders. They found no significant differences at grade 6, but at grade 10 clause length and T-unit length were greater when 'teacher' was the audience than when 'best friend' was the audience. Witte et al (forthcoming) found audience to be a significant factor in scores of freshmen on twelve variations of a single essay topic: as audience specification increased, scores increased. It would appear that the specification of a low intimacy audience requires a writer to demonstrate his maximum syntactic control, and also to use a wider range of strategic competence features than would be used with a more intimate audience.

3.4.4. Mode of discourse

Kincaid (1953) observed a large variation in writing performance from day to day, which was more marked in better than poor writers. He suggests that this may be directly related to varying modes of discourse for essay tasks, i.e., that writers perform differentially on narrative, descriptive, argumentative or expository tasks. Crowhurst and Piche (1979) found a significant variation in the syntactic complexity of writing by mode of discourse for sixth grade and tenth grade children. At both levels, argumentative essays were syntactically more complex than either descriptive or narrative essays. This finding agrees with those of studies by Rosen (1969) and Perron (1977). Galbraith (1980) believes that argumentative and expository writing is more demanding than narrative or descriptive writing:

> *The goal of expression places no constraints on the form of the final product; whereas the goals of coherence and self presentation do. This means that when the latter goals govern the form in which ideas are expressed, there is the possibility that ideas will be distorted. (p.365)*

Pollitt et al (1985) found, similarly, that writing which involves explaining is more difficult than writing which involves describing. Quellmalz, Capell and Chou (1982) studied the writing of twelfth grade students on essay tests and found clear variations in levels of performance on narrative and expository discourse types. Freedman and Pringle (1981) found that 12 year olds were able to realise the conventional schema for story structure most of the time whereas they were only able to realise the conventional schema for argument structure 12.5 % of the time. Freedman and Calfee (1983) found significant variation between scores on compositions requiring opinions and compositions requiring quotation, suggesting that even within the 'expository mode' there are sub-modes or genres which draw upon different aspects of writers' writing ability.

3.4.5. Culture-related expectations

Hoover and Politzer (1980) say that bidialectal students whose dominant dialect is not that of the majority culture in the U.S A. perform disproportionately poorly on composition tests. We saw in Section 3.3.3. that White and Thomas (1980) provide some evidence of this. However, Hoover and Politzer go on to claim that bidialectal students are victimised by negative attitudes to their language and culture on the part of teachers, citing Hoover and Politzer (1977) and Shuy and Fasold (1973). They consider that such attitudes not only affect the teacher's assessment of the student's communication skills, but also interfere with the student's ability to communicate successfully because the student senses the teacher's rejection:

> *The papers you write in class - their whole attitude to them is bad. They say the structure is not too good and the style is bad, when you are writing from your heart ... it's like they are rejecting your whole culture. (p. 198)*

Toelken (1975) set up a culturally sympathetic study to find out why Native Americans drop out of university. The first reason given was almost universally 'Freshman Composition'. Although in some cases this was because of arrested literacy, deeper investigation showed that the Native American students found the composition assignments "illogical" or "impossible to do"; many of the assignments were "not topics at all" or "things you just can't talk about". For example, autobiographical topics were not seen as appropriate for college-age students, since Navajos only think of themselves as 'persons' as a result of age and experience; 'your plans for the future' was seen as non-logical, since Navajos do not take the same long view as Anglo-Americans, or as "tempting fate". Further, the concept of an orderly theme, the concept of logical syllogisms, the notion of a required sequence, the teacher's distaste for repetition, and other task expectations posed problems for these students. Many of these students had English as their first language, but they had the same difficulties as those who were bilingual.

3.4.6. Linguistic characteristics

Harpin (1976) found that for elementary school children small changes in the wording of questions led to considerable changes in topic interpretation. Baker and Henman (1983) give linguistic complexity as one of the features to be considered when designing writing test tasks, but present no data. O'Donnell (1968) investigated 969 SCE 'O' level scripts in physics in order to establish whether a link existed between the language of the questions in terms of syntax, vocabulary, and the combination of these, and the examination performance in terms of the question chosen by the candidate (on the assumption that they choose the questions which seem easiest) and the scores actually attained on the questions chosen. O'Donnell found that the most frequently chosen questions were not the questions which got the highest scores. O'Donnell found that syntactically complex questions were less popular than syntactically simpler questions, but that syntactically complex questions

received higher scores than syntactically simpler ones. He also found that the choices of question made by writers were not a matter of chance, and suggested tentatively that lexical complexity might make a question unpopular. O'Donnell feels that:

> *Every examination presents the candidate with a language task which he must be able to perform in order to undertake the examination - whatever the subject in which he is ostensibly being examined. Nobody seems to have any clear idea about the dimensions of the task in any given case: the language component, in fact, being largely taken for granted. (p.1)*

Brossell and Hoetker Ash (1984) in their study of 21 different essay topics "came away with the feeling that" small syntactic variations do not have "much of an effect on essay exam scores" (p. 145). However, Pollitt et al (1985), in their study of comprehension questions in English and French, suggested that 'content' words (meaning-bearing words like nouns, verb stems and adjectives) were taken at face value at the expense of functional details.

3.4.7. A choice of topic?

Wiseman and Wrigley (1949) investigated the effect of testees' question choices, and found significantly different mean scores for each essay topic. They considered that some of the difference was due to the markers but that most of it was due to the children: 'poor' children chose pedestrian topics they thought 'safe', whereas markers had been told to value fluency, vitality and force and to de-emphasize usage. They concluded that although there were real differences in apparent difficulty level of the topics, these were in fact due to real differences in ability levels of the children. This conclusion seems surprising in view of the fact that they found a larger marker effect than child ability effect in their data analysis, and O'Donnell's study, reported above, suggests otherwise.

Coffman (1971b) particularly stresses the inadvisability of offering a choice of topic when testees come from widely differing backgrounds. Diederich (1971) suggests six papers on different topics, but Jacobs et al (1981) believe that two pieces of writing on different topics are enough, as long as each testee answers the same two questions, and provided that the topics have been pretested to ensure that they do not produce significantly different performance. Britton et al (1966) argue that any topic "may be anathema" to some candidates and ideal for others (p.3). They favour a choice of topic, but recognise that this introduces different variables, because it is exceedingly difficult to construct two equally difficult essay questions, and because students are not good at choosing the topic with which they will do best. Meyer (1939) considered the arguments on both sides: a choice of questions allows students a chance to write about a subject they know about, and is a way of remedying limited sampling; on the other hand, a choice implies that students are able to decide:

> *...first, what the average score on the whole test is going to be, and secondly, what the average score on each question will be. This would involve, among other things, knowing the real difficulty of the questions; knowing the scoring standards; and knowing who is to do the grading. (p.155)*

Meyer recommended that essay tests with a choice of questions should be discontinued. Poetker (1977) gives the same advice, suggesting that allowing students a choice does not work to their benefit, a suggestion supported by the findings of O'Donnell (op cit). Pollitt and Hutchinson (n.d.) looked at the relative attractiveness of English essay topics for candidates at the top end and the bottom end of their sample, and found that the mean score for candidates at the top end matched the rank popularity of the essay topics, whereas the opposite was the case for candidates at the bottom of the sample. They suggest that the ability to choose a topic favours strong candidates over weak candidates.

Permitting testees a choice of topic introduces a range of additional variables to the assessment: it becomes essential to discover whether observed differences in scores are due to real variations in writing proficiency, or to variations in the topics. It is also much more difficult for readers to rate reliably when not all papers are on the same topic (Coffman, 1971(b); Diederich, 1974). The problem with not allowing a choice is to ensure that the content of the topic, however constrained, is equally within the range of every writer. Hilgers (1982) recommends providing the information necessary to ensure that all writers have equal familiarity with the subject in order to uniformly interpret the topic. However, the linking of the task to a reading or listening text brings in comprehension variables; even a wholly visual input is not problem-free. Any input material must be well within the range of ability of all testees in the skills it demands, and time to process the input material needs to be included in the total test time (but not the writing time).

3.5. What makes questions difficult?

Text structure research (e.g., Kozminsky, 1977; Meyer, 1975, 1977; Meyer & Rice, 1982) and document design research (e.g., Swartz et al, 1980; Wright, 1981a, 1984) indicate the importance of the quality of the input for comprehension and production. Although this research has not looked specifically at essay titles, Kozminsky (op cit) looked at the effect of alterations in titles to texts and found that comprehension and interpretation of the text was affected by changes in the propositional structure of titles. Swartz et al (op cit) focussed specifically on headings in technical documents, and found that clear headings led to better prediction of text content, and higher success rates at matching headings with their texts. Hinsley et al (1977) applied 'miscue' research (e.g., Rumelhart, 1975) to algebra problem-solving tasks and found that subjects categorized problems on early verbal cues after hearing one-fifth of the text: they paid less attention, and some

misheard, the rest of the problem, so that they perceived the whole problem in accordance with the miscue they had been fed at the outset.

Research specifically into issues of question difficulty for essay questions has begun only recently and is still in a predominantly descriptive stage. Rosen (1969), accepting that essay questions vary in difficulty, proposed the following criteria to be considered when designing questions:

1. *awareness of the implications of the question statement (type of writing demanded)*
2. *linguistic characteristics*
   - a) *avoidance of metalanguage*
   - b) *explicitness*
   - c) *semantic consciousness*
3. *area of experience to be drawn on*
   - a) *personal/impersonal*
   - b) *highly charged emotionally/not*
4. *psychological*
   - a) *appropriacy for age group and other known affective factors*
   - b) *response not pre-empted*

Coffman (1971b) points out that the more complex the structure of the question, the more time testees need to think about and compose a response, and also the more possibility that the testee will misunderstand the question. Poetker (1977) and the New York State Education Dept Bureau of Social Studies (no date) have general suggestions for constructing essay questions, but like Coffman do not present any data. Greenberg (1981) investigated the effects of four different kinds of essay questions: high cognitive demand; low cognitive

demand; high experiential demand; low experiential demand. She found that varying the questions along these dimensions produced no significant differences in writing performance. Brossell (1983) used six tasks each with high, moderate and low rhetorical specification versions. He found that the versions with moderate rhetorical specification yielded the highest mean scores (a 'moderate' rhetorical specification is described as a short introductory statement then an instruction to respond). Baker and Henman (1983) suggest the elements of a task structure approach.

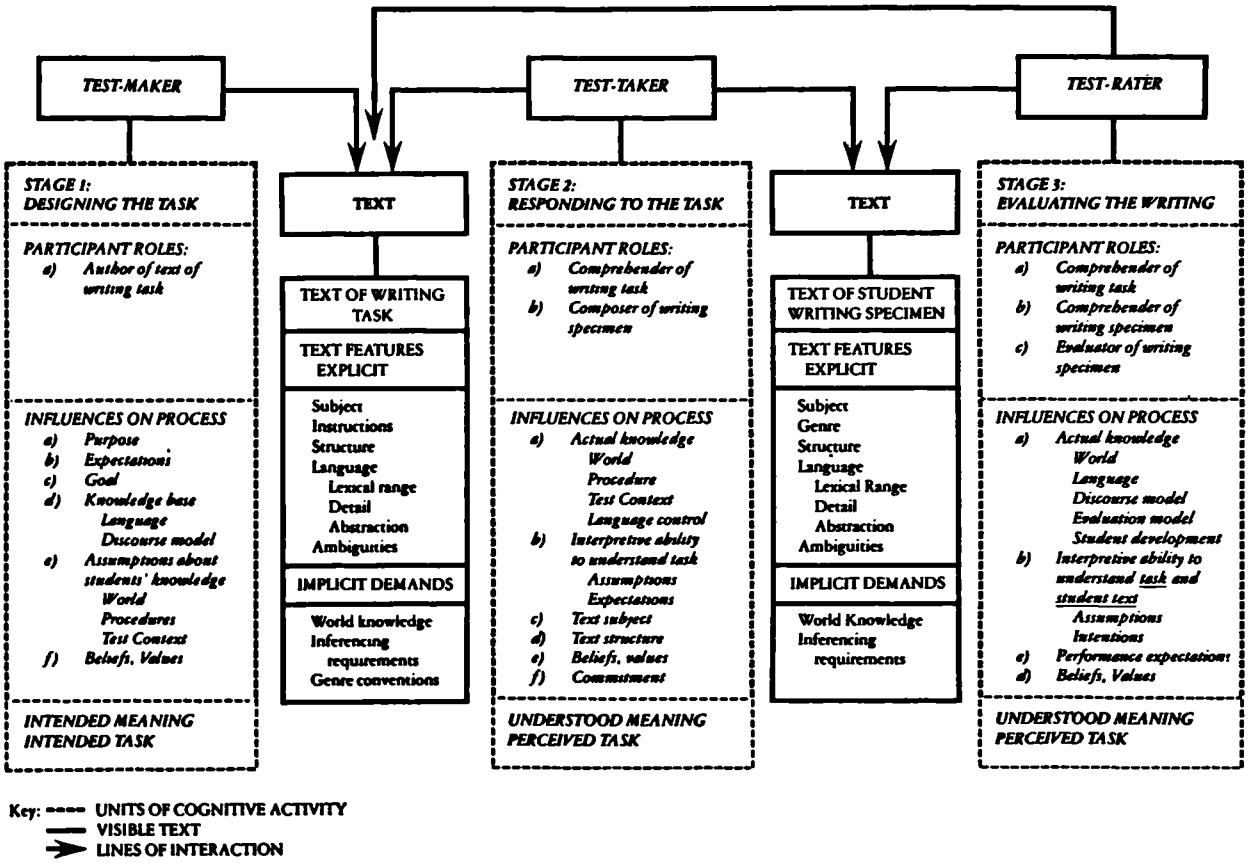
> *task description (i.e., outcome statements)*
> *content limits (i.e., specification of what should be included)*
> *linguistic features/complexity*
> *cognitive complexity*
> *format (of item and required response)*

Meredith and Williams (1984) suggest four considerations for design of an essay 'prompt': appropriacy for testees; breadth; freedom from creation of emotional responses; consistency with the purposes of the assessment. They do not, however, provide empirical data or practical guidelines. Ruth and Murphy (1984) report that in the Bay Area Writing Assessment Project their attempts to understand and describe the properties of tasks for writing assessments led them to constructivist theories of reading comprehension. They consider that:

> *... the 'meaning potential' of any given task is relative to the linguistic, cognitive, and social reverberations set off in the respondents. Both the language of the topic and the general knowledge of the participants interact in a writing test to determine what meanings the topic may elicit. (p. 413)*

In such a view of question difficulty, the difficulty resides almost wholly in the writer and little in the question itself. They see the process of designing a writing task as almost wholly a matter of accounting for human interactions :

FIGURE 2.3.2.: PARTICIPANTS, PROCESSES AND TEXTS IN A WRITING ASSESSMENT EPISODE



This is not a view shared by Pollitt et al (1985), who consider it possible to identify and predict sources of difficulty in questions. After studying questions and responses to them in five subject areas, Pollitt et al conclude that question difficulty is associated with three separable facets of the task, within each of which it is possible to identify several specific causes of difficulty:

Subject or concept difficulty
degree of familiarity
abstractness of mode
abstractness of idea

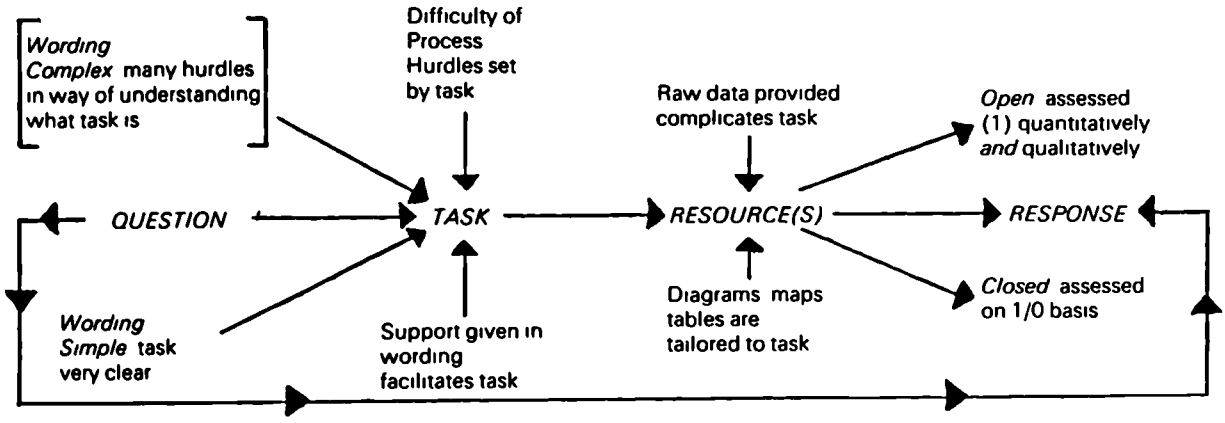
Process difficulty
explaining
generalising from data

selection of data relevant to general theme
identifying a principle from data
applying a principle to new data
forming a strategy
composing an answer
cumulative difficulty
need for monitoring logical consistency

Stimulus difficulty
open/closed response
leaders, cues, cluse
tailoring of resources
provision of answer structure

Pollitt et al propose the following model for the construction of an essay question:

FIGURE 2.3.3.: CONSTRUCTING A QUESTION



White (1986) reports on a large-scale assessment of the writing of children aged 11-16 conducted by the APU (Assessment of Performance Unit of the Department of Education and Science of England and Wales). Tasks in this study had been developed according to an earlier APU task framework (Gorman et al, 1981), in which there were four task dimensions.

> *writing type (narrative/ descriptive vs reflective/ analytic)*
> *degree of control ('outcome space')*
> *source of subject matter (experience vs learned knowledge)*
> *range of purposes*

White comments:

> *One of the results of using a range of different writing tasks ... is that we have been alerted to the task-specific characteristics entailed in children's ability to write purposefully and well. Ease or difficulty in writing is very much a feature of specific tasks ... (p. 25)*

Currently there is some interest in the use of item response theoretic methods, usually Rasch analysis, to achieve a statement of the difficulty of essay test items which is free of person-ability (Rentz, 1984; Henning and Davidson, 1986; Pollitt, personal communication). As we saw in section 2, the issue here is that of the stochastic independence assumption of latent trait methods, and discussion in this area is continuing.

4. TESTING WRITING IN A SECOND/FOREIGN LANGUAGE

Clearly, the principles for the testing of writing remain the same for the testing of the writing of second language writers. There is little research evidence to indicate whether or not findings from studies of English L1 writers and their writing on essay tests can be applied to ESL writers writing essay tests in English. Throughout this chapter,

however, insights from first language studies and from second/foreign studies have been interwoven, and this researcher has become increasingly convinced that not only are there no contradictions in the issues, methods or results from the two types of studies, but that this interweaving provides a much fuller picture than would be possible if only the research from one or the other were surveyed. This section therefore serves only to highlight some particularly crucial considerations for the testing of the writing of non-native writers.

4.1. Reader variables

There appears to be little research which focuses on the scoring behaviour of raters of first language writing by comparison with the scoring behaviour of second/foreign language writing. Carlson et al (1985) report that in their study the mean scores of ESL readers and of "regular English teacher readers" were nearly identical. They believe:

> *... the careful training procedures employed in this study were sufficient to overcome any differences in rating strategies between the two types of readers that might otherwise have occurred. (p. 61)*

In contrast, Robinson (1985) found a significant effect for rater background (ESL/non-ESL) on the characteristics of student essays which were most valued and on the scores assigned, despite careful training.

4.2. Procedural variables

There have not to date been any suggestions that the writing of ESL writers should be scored using any procedures different from those used for the scoring of the writing of English first language writers. Purves (1984) comments on the difficulty of finding a procedure for scoring compositions which can be applied across languages, including non-alphabetic languages, but does not suggest that ESL and non-ESL writers writing in English cannot be judged by the same procedure. As we saw in

Chapter 1, section 1, recent developments in thinking about the composing process in the first language have been paralleled in second language composing research. The current characterisation of the construct of 'writing' or 'composing' has made attempts at the objective testing of writing inappropriate and frequency count procedures unpopular. We may see the debate at the moment as centering on whether holistic or analytic procedures are more appropriate, although it would be unwise to expect that such a debate will ever reach a conclusion.

4.3. Writer variables

A study by Breland and Jones (1982) showed that although for both English L1 and ESL writers discourse level characteristics were the best predictors of holistic essay scores, for the ESL group syntax and lexis scores were relatively more important than for the English L1 group.

Fein (1980) investigated English L1 and ESL writers from 'equivalent' college courses to discover whether systematic differences could be found in their writing, and whether they received significantly different scores on a holistic evaluation of the same task. He found that although the ESL students scored significantly lower on a holistic assessment and on an error count, analysis of their content, organization or style showed little difference from the English L1 students. Detailed analysis showed that ESL writers consistently had more errors than English L1 writers with the same score: therefore, whatever criteria the judges were using, accuracy was not their only one. He also found that weak English L1 writers consistently generated fewer ideas and were more irrelevant than the ESL writers, and hypothesized that organization and content were compensating in the ESL writers' scores. Fein suggests that the differences may be due to language acquisition at work, and that fluency and discourse level characteristics are acquired before grammatical accuracy, or else that they are due to "fossilization of error".

It was first suggested over twenty years ago (Kaplan, 1966) that rhetorical patterns differ from culture to culture: that while English expository writing has a 'linear' development, writing in other cultures develops in different rhetorical ways. Kaplan believes that these culturally determined rhetorical differences lead to breakdown in communication between the second language writer and the first language reader:

> *Instructors have written on the writing efforts of (foreign) students, comments like: "The material is all here, but it seems somehow out of focus," or "Lacks organization," or "Lacks cohesion". And these comments are essentially accurate. The student's perception appears out of focus because the student is employing a rhetoric and a sequence of thought which violate the expectations of the native speaker. (1972, p.8-9)*

Kaplan suggested some representations of the discourse patterns of various cultures: Japanese, for example, is typically pictured as developing through a spirally rhetorical structure, circling the subject but not approaching it directly (Kaplan, 1982; Onaka, 1984). Research into the rhetoric of, for example, Arabic (Thompson-Panoŝ & Tomas-Ruziç, 1983; Koch, 1984), Chinese (Tsao, 1983), Farsi (Dehghanpisheh, 1973), French (Regent, 1985), German (Clyne, 1981), Greek (Tannen, 1979, 1980), Japanese (Hinds, 1983; Kobayashi, 1984), and Spanish (Santiago, 1968; Santana-Seda, 1974) has consistently supported the view of cultural differences between the rhetorical structures of written discourse in different languages.

On the other hand, there have been objections to this discourse-level version of the Sapir-Whorf Hypothesis. James (1980) has described Kaplan's designation of English expository prose as having a 'linear' rhetoric as ethnocentric. Das (1985) reported a study in which the writing of bilingual students in English and their L1 was found to be equally deficient by a panel of judges. Moragne è Silva (1986) presented a case study of an English/Spanish bilingual writer who showed similar

deficiencies in both languages. Mohan & Lo (1985) criticize Kaplan's claim of interference in the English writing of Chinese ESL students from the organization patterns of Chinese expository prose on the grounds that such claims should be supported by 1) contrastive analysis of L1 and L2; 2) error analysis of the L2 learner's performance in L2; 3) clear evidence that the errors found are due to transfer: they find Kaplan's work lacking in these areas.

Certainly it is the case, as Houghton & Hoey (1983) point out, that "in general linguists are not yet in a strong enough position to be able to support without reservations the notion of contrastive rhetoric." (9) Kaplan's contribution has been primarily in model-building rather than in rigorous scientific proof, a contribution which, as Houghton & Hoey also point out, is both essential and eminently academically acceptable (9-10). Responding to Mohan & Lo (op cit), Ricento (1986) states: "Few scholars working in the area of contrastive rhetoric would disagree with Mohan & Lo's claim that it is risky to infer L1 rhetorical patterns from ESL student compositions." (565) However, he goes on to urge caution, since Mohan & Lo's own claims are "made on the basis of a small corpus of short (Chinese) texts which happen to utilize certain (characteristics of) English expository prose" (ibid). In Mohan's (1986) follow-up to the Mohan & Lo/Ricento exchange, he states: "Assumptions about similarities and differences in cross-cultural discourse studies should be justified as far as possible and stated as clearly as possible as working hypotheses." (572)

It has often been argued that non-native writers of English can be expected to have difficulty with the rhetoric of English, to a greater or lesser extent depending on their own culture's rhetorical differences from English. We saw in section 3 that bidialectal and bilingual writers using English encounter many problems and frustrations, some of them before they even begin to attempt to write. Clyne (1981), who studied

German speaking migrants in Australia, identifies some of these cultural/rhetorical differences:

> *... culturally conditioned formalism...rules for the writing of academic treatises and of essays in non-language subjects within the education system, where presentation may override the knowledge which is the object of the essay. Such rules are difficult for people from other cultures to understand, let alone adhere to. ...Adolescents and mature age students who received much of their education in a non-English-speaking country often fail in Australian schools and universities in spite of adequate knowledge of the appropriate subject and a high competence in it, because they have not been sufficiently trained to abide by formal rules which reflect features of a culture of which they are not part. (p.62)*

Clyne describes features of German academic discourse behaviour which are quite unlike those of 'Anglo-Celtic', and which can explain characteristics of German L1 writers writing academic discourse in English. Given the linguistic closeness of German and English, we might expect that such differences would be more rather than less pronounced for other cultures. A great deal of research remains to be done in this area.

4.4. Task variables

There are no studies to date of the difference in effect of topic choice on ESL writers by comparison to English first language writers. It would seem especially important with non-native writers of the language to select topics which are free of cultural bias in content and schemata. Topics should not relate to some aspect of British, or English-speaking, culture which not every testee can be sure to know of. When the test is administered to testees in their own countries rather than in Britain after some period of residence, this is even more important. Topics which would be culturally or politically sensitive to some students, or which in other ways presuppose a shared set of values or shared schemata, are best avoided, unless it is possible to be sure that these

qualities will be shared among all cultures. In the context of this study, academic topics are appropriate and can avoid the problems of bias on many of the grounds discussed above: the question of bias due to greater or lesser familiarity with the academic knowledge required for the topic of course arises, but since this approaches one of the main research questions of the study it will not be discussed here.

We have seen (Sections 3.4.2. and 3.4.3.) that both the audience variable and the mode of discourse variable exert considerable influence on writing performance. A study of EFL writers by Arena (1975) suggested that their proficiency in writing in the narrative or descriptive modes does not carry over into their expository writing. It seems unlikely that audience and mode of discourse variables operate any differently for second/foreign language writers than for first language writers, but there is to date little evidence in this area. In both cases the test constructor will want to select an audience and mode of discourse which are valid in terms of what the testees will be expected to do with writing in the context to which the test related, and which also are of an appropriate difficulty level for the testees and context.

We can expect that all the points made in the discussion of linguistic structure and complexity from the point of view of first language writing (Section 3.4.6.) would apply to second/foreign language writers also. There is at present little available research into this question. Hirokawa and Swales (1986) investigated the effect of two different levels of formality, 'simple' and 'academic', in essay questions on the scores assigned to ESL writers, and found no differences in scores on essays of the two types, although there were statistically significant differences in several features of the writing: compositions on the simple topics were longer, contained more subordination, more use of first person singular, and more morphological errors; compositions on the academic topics had a higher proportion of Graeco-Latin lexis, had fewer syntactic errors and a smaller total number of errors.

4.5. Task design

It was seen in section 3.4. that there has been little work in task design for writing tests generally, and what there is has been very recent. There are no available research studies of second/foreign language writing tests which take the variables described in 3.4.1. to 3.4.6., investigate them either theoretically or empirically, and apply the results to task design for second/foreign language writing tests. We are in the position of having no equivalent for Section 3.5. Clearly, however, the more we see a writing test as an almost mystical encounter between a writer and a reader, as Ruth and Murphy (forthcoming) seem to do, the more difficult it will be to establish meaningful parameters for task design, and the difficulty is surely exacerbated when the encounter is a cross-cultural one. But the more we accept a social constructionist view of the writing and reading processes, the easier it will be to place the writer and the reader within an interpretive community. Task design then becomes a matter of defining the interpretive community within a discourse community, and discourse communities may exist across cultures, and in fact do exist across cultures, in business, the professions, and research specialisations.

5. OVERVIEW

It is now possible to distil from the work surveyed in this chapter an overview of some important characteristics which a well-designed test of the writing of non-native English speaker applicants to postgraduate course at British universities and colleges should possess.

> 1. *The scoring procedure should be carefully developed to reflect the important characteristics, in terms of criteria and standards, of proficient writing for the context. A protocol of linguistic, rhetorical and communicative criteria should be provided to inform the scoring. To the extent that the scoring of any question is content-related, a content protocol should also be provided stating the content criteria*

for this question, i.e., proposition, main ideas, supporting detials, some examples, and the relationships between the content levels and items

2. *Raters should be well-trained and training should be refreshed often; the raters' backgrounds should be taken into account when selecting raters. Raters should receive very clear guidance about the value to be ascribed to such features of answers as length, spelling, and non-standard linguistic and rhetorical features.*

3. *Multiple-marking should be used whenever possible.*

4. *Tests should consist of at least two compulsory topics which will be given equal weighting in the final score, unless valid reasons can be presented for some other weighting.*

5. *Tasks on the writing test should be communicative; that is, they should be placed within a realistic and meaningful context and should state a purpose and an audience. The evaluation criteria for the tasks should also emphasize the communicative nature of writing more than accuracy.*

6. *Tasks should balance freedom and constraint, offering a reasonable 'outcome space': writers, including non-native writers, need freedom to value and respond to a task in a personal way; at the same time it is necessary to provide topic, audience and mode of discourse constraints in order to ensure score reliability.*

7. *Tasks should be fair and reasonable; that is, each topic should be equally accessible as regards its knowledge requirements to all writers likely to be required to answer the question; and no task should be more difficult nor more demandingly graded than a similar real-life task would be for native speakers.*

We are not able, as a result of the research reported here, to specify the specific characteristics which should apply to a test of second langugae writing in academic settings: this problem will be approached through the investigations of the constructs of 'language proficiency' and English for Academic Purposes/English for Specific Purpose, reported in Chapter 3.

CHAPTER THREE

THE PROBLEM IN CONTEXT (2): RELATING VIEWS OF PROFICIENCY TO SPECIFIC PURPOSES

INTRODUCTION

In the development of a framework within which we can investigate the testing of second language writing in academic settings, it has so far been necessary to give attention to the fields of writing, language testing, and the testing of writing. But this study is not simply about the testing of writing, but about the testing of writing in academic settings, that is, academic writing, and more than this, the testing of academic writing for specific purposes. This construct must be placed in the context of the debate of the past ten years over the construct of language proficiency. The debate is of course relevant to a conceptualization of first language proficiency also, but it has been particularly critical for researchers into second/foreign language testing. Language programmes have been designed and implemented, and language tests have been developed, used and interpreted, based on a clear set of assumptions about the nature and structure of language proficiency.

Davies (1981: p.182) has given the view that "General Language Proficiency ... is essentially a non-issue theoretically. At the same time, the practical implications are important." Underlying specific purpose testing and academic purpose testing is clearly a construct of general language proficiency as at least partially divisible; the evidence for and against this construct will be reviewed in the first section of this chapter. The second section of this chapter looks at the concepts 'specific purposes' and 'academic purposes', focussing on the testing of each. Then these two approaches are brought together , to a focus on

writing in academic settings in the third section, and to a focus on testing writing in academic setting in the fourth.

1. GENERAL LANGUAGE PROFICIENCY

An understanding of the concept of language proficiency is vital to how we view the nature and demands of language testing. Farhady (1982. p.44) believes that "language proficiency is one of the most poorly defined concepts in the field of language testing". It is particularly vital when the concern, as in this study, is not with general tests but with 'specific' tests.

1.1. Views of proficiency

'Proficiency' is one of a number of terms in foreign language teaching and testing which has been used a-theoretically or pre-theoretically over a period of time, and which has gone through shifts in its connotations to reflect more general paradigm shifts in attitudes and approaches to foreign language teaching and testing. It is common at present to think of proficiency as similar to, or even synonymous with, 'performance' in the competence/performance dichotomy originally proposed by Chomsky (1965) and referred to by Canale and Swain (1980) as communicative competence and communicative performance (Richards, 1985). Whereas competence refers to what is known about the rules of the language and of using it , performance (and therefore proficiency) refers to the language user and what she or he can do with the language. The distinction here is the same as that made by Widdowson (1978) between usage (competence) and use (performance). This view of proficiency means that proficiency is always understood with reference to specific situations, needs, purposes and problems: in this sense proficiency is always specific, always for something. Spolsky (1973) talks of "knowing a language" and points out that a layman would be likely to make a functional statement if asked whether he 'knew' a language, for example, " I know enough French

to read a newspaper" or "He can't speak enough English to ask the time of day" (p.166). Proficiency also seems to be used to refer to the degree of skill with which the language user can do something, as the examples above show. From the foregoing it is clear that proficiency, in this view, is not divided up into discrete 'bits', but that when language users use the language to perform some real-life task they call on and apply several or many aspects of their language competence at the same time, integrating what is needed to successfully use the language for their present purpose.

1.2. Theories of Proficiency

In contrast to the pre-theoretic or a-theoretic views of proficiency, theories of proficiency relate it more closely to competence, to an underlying knowledge or set of knowledges. Vollmer (1981) states that a large number of researchers have started from an assumption that there are a number of underlying competencies, and have concentrated on identifying and naming those competencies that could be related to language behaviour on the performance level (p. 154). He cannot see that general language proficiency has any place within the framework built up by J.B. Carroll (Vollmer cites Carroll, 1961) and others.

Theories of language proficiency are closely related to theories of intelligence. Spearman (1904) suggested the existence of 'g' - general intelligence. Thurstone (1938) argued that there are three factors of mental ability: M (memory or rote learning), V (verbal relations) and W (word fluency). Verbal abilities and intellectual abilities were seen as bound up together. J.B. Carroll (1941) investigated verbal abilities through factor analysis, administering 42 tests to the same candidates. He identified nine factors: his principal factor, C, he suggested corresponded closely with Thurstone's V, but suggested that his data showed V as being actually three factors: C, J and perhaps G. Thurstone had described V as an ability to manipulate ideas in discourse, but

Carroll described his C as knowledge of verbal tokens underlying manipulation of ideas and relationships; J is more like V, being described as reasoning ability, or ability to handle verbal relationships. His G showed loadings from so many tests he could not attempt to characterize it, but it was particularly well-represented by the Handwriting Speed test. Thurstone's second factor seemed to be represented by two factors in Carroll's study, A (the speed of word association in restricted contexts) and E (the rate of production at discourse level). The other factors were: B (rote learning - Thurstone's M); D (speed of articulation); F (speaking ability); H (speed of attaching verbal response to stimulus).

Carroll had hoped to discover whether 'general speech fluency' is "an operational unity unrelated to intellectual abilities" (op cit, p. 281); he concluded that his C involved "some sort of intellectual verbal ability" (p. 293), and went on to say that:

> *...this factor represents the individual differences in some aspect of the ability to learn various conventional linguistic responses and to retain them over long periods of time. The factor represents differences in the stock of linguistic responses possessed by the individual - the wealth of the individual's past experience and training in the English language. (loc cit)*

Thus Carroll's first factor was at once a language factor and an intelligence factor. The possibility of a close relationship between language and intelligence was thereafter somewhat unexplored for some time, a time during which the accepted view of language proficiency became that of Lado (1961) and Carroll (1961), in which the language system was divided into skills (listening, speaking, reading, writing) and elements (pronunciation, syntax, lexis, and what Lado called 'cultural meanings'; other elements were proposed by other researchers).

It was only when Oller began to develop his notion of 'expectancy grammar' (e.g., 1974), a single internalized grammar upon which all

language ability is based, and to claim that all language tests are essentially measuring the same thing (1978a), that the view held by Lado and many others was seriously questioned, and having been questioned was seriously explored by its proponents. Oller was motivated to explore this notion by his observation that language tests tend to correlate at around .7 with each other; he further noted that 'language' tests and 'intelligence' tests tend to correlate at the same level, and suggested that they are actually measures of the same thing (op cit). He referred to research (Stump, 1978) that showed that variability in language proficiency accounts for the majority of the variance in measures of 'intelligence'.

Oller (1979a) described three hypotheses which might account for the structure of language proficiency and which would be amenable to empirical proof. Oller used the term 'competence' in referring to each of his three competing hypotheses, thus making language proficiency a knowledge phenomenon rather than a skill phenomenon, and placing it solidly in the theoretical rather than the practical sphere. His first hypothesis was the Divisibility Hypothesis:

> *...there will be reliable variance shared by tests that assess the same component, skill, aspect, or element of language proficiency, but essentially no common variance across tests of different components, skills, aspects, or elements (p. 425)*

Second, the Indivisibility, or Unitary Competence, Hypothesis. This view of language proficiency posits that:

> *...there will be reliable variance shared by all of the tests and essentially no unique variance shared by tests that purport to measure a particular skill, component, or aspect of language proficiency. (loc cit)*

Third, the Partial Divisibility Hypothesis:

> *...there will be a large chunk of reliable variance shared by all of the tests, plus small amounts of reliable variance shared by only some of the tests. (loc cit)*

Oller and Khan (1980) suggest that the Divisibility Hypothesis was refuted by work done by Valette (1964), Darnell (1968), Oller (1971), Oller and Conrad (1971), Oller (1972) and " a flurry of testing research worldwide eventuating in many replications of the basic findings" (p 4). These basic findings were that a single test was about as good a measure of overall proficiency as a more complex battery of tests. Oller and Khan felt in 1980 that :

> *...at present there seem to be two possibilities: either the general factor of language proficiency accounts for all of the reliable variance in mental tests or nearly all of it. There can be no doubt any longer that such a general factor exists and is best explained as a language factor (p. 5)*

However, this claim was tempered somewhat in the conclusions.

> *Does all of the foregoing prove that there is only one factor of language proficiency and that it is in fact indivisible as certain pragmatic theories might lead us to suppose? Certainly not. ...What is demonstrated is something a bit weaker: that the global language factor is almost certainly the most important element (perhaps the only element) in many tests where it might not have been expected to hold sway. (op cit, p. 8)*

Oller and Hinofotis (1980) compared their search for empirical validation of the unitary language competence hypothesis with Spearman's (1904) argument for a general factor of intelligence, and reasoned from this that the statistical method used in the investigation of general intelligence could be applied to the indivisibility/divisibility question. The factor analytic method they used was "factoring a variety of language tests to a principal components solution and then testing for a general factor by using the loadings on the first principal component to predict

the original correlation matrix" (op cit, p. 15). This method has since been criticised by a number of researchers.

Vollmer (1981), first points out that the principal factor model is superior to the model used by Oller and Hinofotis, and then reminds us that all classical forms of factor analysis are explorative, that is, "they work even without any piece of foregoing theory" (p. 167). He continues:

> *We will never be able to select the meaningful factors from those that are pure artefacts. In other words, the structural hypothesis of a unitary factor, being the simplest under conditions given, has always quite a good chance of being confirmed, even if it does not represent at all any adequate decription of the relationships among the several linguistic skills. (loc cit)*

In this sense the choice of one model over the other is always a matter of personal choice: Vollmer and Sang (1983) quote from a study by Scholz et al (1980), which Oller has often cited in support of the indivisibility hypothesis, to prove their point:

> *...the problem was to choose between the multiple-factor solution (the varimax rotation) and the single-factor solution (the first factor of the principal components analysis). Choosing the latter would eliminate the divisible competence hypothesis, and choosing the former would eliminate the unitary competence hypothesis (Vollmer and Sang, op cit, p. 64)*

Vollmer (op cit) recommended the use of confirmatory factor analysis, which permits a statistical comparison between theory-guided structural predictions and test results. Palmer and Bachman (1981) reported a confirmatory factor analysis study which they felt "found strong evidence supporting Oller's divisible language competence model", i.e. disconfirming Oller's original position (p.144). Hughes criticises the use of some of the studies cited to support the unitary competence hypothesis: he suggests that some of the tests used were not actually measuring what

they were supposed to be measuring, and therefore it is not surprising that they do not load on the same 'unique' factor, and, since they typically require testees to identify the appropriate written response it is not surprising either that they should correlate quite highly with reading tests (1981: p.234). Hughes also puts forward another criticism: that heterogeneous groups of testees leads to false unifactorial solutions. Too great a range of ability will appear to favour the indivisibility hypothesis because individual differences are obscured; he cites evidence from Oller and Hinofotis' own study (op cit) and from a study by Yorozuya and Oller (1980).

More recently (1983) Oller has accepted that "the strongest form of the unitary hypothesis was wrong" (p. 352). He accepts that the weight of empirical evidence, including reanalyses of data originally used in defense of the indivisibility hypothesis, and the theoretical arguments put forward, make that position untenable. However, Oller maintains that there must be "a general factor underlying performance on many language processing tasks" (op cit, p. 353). Thus the work of J.B. Carroll in 1941, described above, in which he found that his factor G loaded on many of his 42 tests, provides partial support for Oller's new position; however, Carroll's 1941 study also found distinct factors across tests. thus the Carroll study can best be described as supporting a partial divisibility model.

Alderson (1981) feels that there is a problem of "level of abstraction or generalisation in the identification or acceptance of the existence of one general language proficiency factor" (p. 187). At the most general level, since language is an identifying characteristic of humans, there must be a general language factor. But at a less abstract level it is clear that different individuals have different skills and different levels of the same skills. Like Davies (1981), he suggests that the reasons why applied linguists are interested in the nature of language proficiency relate to the practical implications for teaching and testing.

1.3. The present position

Oller's arguments in favour of the unitary competence hypothesis rested heavily on specific choices and uses of factor analysis: his opponents' arguments against the hypothesis similarly rested, for a number of years, primarily on alternate choices and uses of factor analysis. Vollmer and Sang (op cit) feel that "the use of factor analysis as a means purely for exploration does not add up to a theoretical understanding, clarification, or even the unification of ideas about foreign language ability" (p 73) and argue for more theory-driven studies. Farhady (1983) provides not only data analysis and re-analysis, but also clearly shows the relationship between the structure of a theory of proficiency and the mathematical manipulations which can appropriately be used to investigate the theory. Farhady suggests that there may well be a 'general' factor, that is, a factor which accounts for a large amount of variance across language tasks, but that such a factor will not exhaust all the reliable variance. Farhady favours a theory of proficiency as composed of a general factor and a number of specific factors. Upshur and Homburg (1983) suggest that such specific factors are not static but are a feature of the language learner at a certain stage of development, changing as the learner develops.

The emerging consensus seems to be that of Farhady; a view of language proficiency as having generality and specificity, although how these are composed and why remains poorly understood.

Oller now (1983) seems to have reached some agreement with his former opponents:

> *I am inclined to agree with the suggestion of Vollmer and Sang (Chapter 3, this volume) and Carroll (Chapter 4, this volume) that hierarchical models of mental abilities may work better than some of the simpler models that have recently been under investigation. (p. 354-5)*

A great deal of work remains to be done in this area, both in theory-building and in experimental research. The small scale case study approach suggested by Vollmer and Sang (op cit) seems to be a fruitful one; until further data is available, however, it would seem that teaching and testing approaches based on constructs of partially divisible proficiency remain defensible, though each attempt to divide up proficiency for teaching or testing purposes will need its own validation. As far as language proficiency testing is concerned, it would seem that at the moment we can do no better than to say, with Vollmer (1981):

> *.. language proficiency is what language tests measure. This circular statement is about all one can firmly say when asked to define the concept of proficiency to date. This is even more so when it comes to the construct of overall language proficiency... (p. 152)*

1.4. Theories of proficiency, and the testing of writing

In relating the foregoing section to the consideration of writing tests, it would appear that there is at present little evidence to suggest that the construction of tests of 'separate' skills should be discontinued, but also a great deal of work to be done to show positively why they should be continued. The issue of whether there <u>is</u> a construct of writing distinct from other language was not treated in Chapter 1: the assumption was made that there was. The discussion of the history of writing assessment and of the backwash from direct writing tests in contrast to standardised testing in Chapter 2 is relevant at this point, to remind us that there are powerful arguments for writing to be separately taught and tested regardless of whether or not a clearly separate construct can be empirically established.

But to say that there are multiple constructs of academic writing which can be distinguished from the global construct 'writing', and which should

be tested separately, is to move into more uncertain territory, as we shall see in the ensuing sections.

2. ENGLISH FOR SPECIFIC PURPOSES

2.1. 'Specific purposes' distinguished from 'academic purposes'

The term 'English for specific purposes' first became well-known in the field of English language teaching. More specifically, English for specific purposes (ESP: originally 'English for special purposes') became "one of the most prestigious fashions" (Robinson, 1980: p.1) of the 1970s in the field of teaching English to speakers of other languages. Strevens (1977a) described, first, a move in second/foreign langauge teaching away from an emphasis on the literature and culture of the speakers of the language and towards teaching for practical command of the language, and second, a move towards the view that the teaching of a language should be matched to the needs and purposes of the language learner.

Carver (1983) tells us:

> *... in reality there is no such thing as English without a purpose, or English for general purposes ... a teaching methodology which includes purpose and specificity in its basic approach is thereby the richer. In this sense, all English teaching is teaching of ESP. (p.132)*

Mackay and Mountford (1978) characterize ESP as "the teaching of English for a clearly utilitarian purpose" (p.2), as being closely associated with teaching adult (post-secondary school) learners, and with English in an auxiliary role. Carver (op cit) refers to a "purpose-related orientation" (p. 134); while he does not refer specifically to adult learners, his emphasis on self-direction and learner-centredness implies mature learners. Carver sees self-direction as taking two forms: learners should

make decisions about when, what and how to study; teachers must systematically attempt to teach learners how to learn.

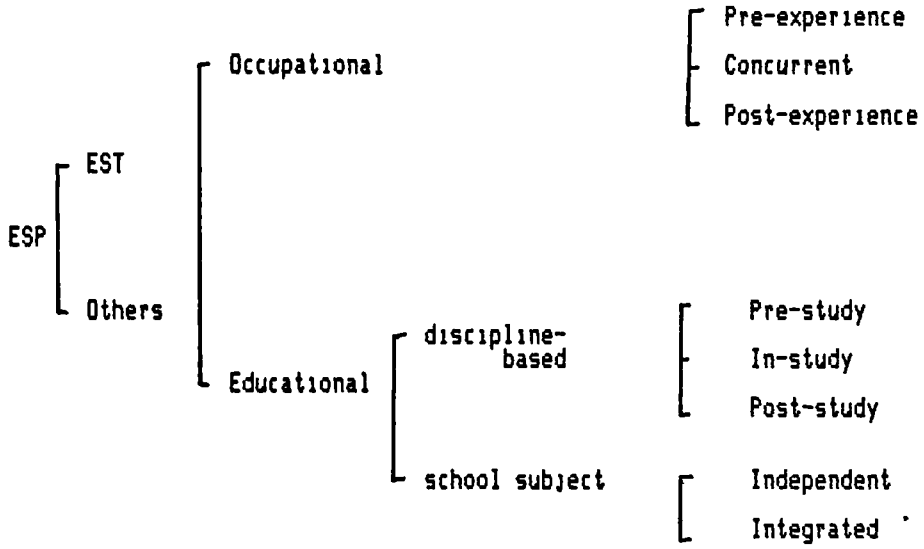
ESP was initially referred to as 'English for special purposes', and closely associated with 'special languages' in the sense of a clearly identifiable register of formal linguistic properties, lexical items, collocations and sentence structures. Mackay and Mountford (op cit), however, prefer to see ESP as an approach to data which represents particular modes of language use that characterize particular scientific, occupational or vocational fields. They place the emphasis on the purpose for which the learner is learning the language and not on the language he is learning. Robinson (op cit) concurs in this distinction.

Researchers in the sociology of education (e g , Halsey and Trow 1971; Bailey, 1977) commonly portray the university as a community, and it is becoming increasingly common to talk and write of 'disciplinary cultures' (e.g., Light, 1974; Becher, 1981). Becher (1986) speaks of the debate over whether academics should be regarded as members of a single profession or whether 'the profession' is more accurately considered as a large number of different professions, tending personally to the latter view.

There have been a number of attempts at classificatory systems for academic disciplines, among the best-known of which are Biglan (1973), who classified disciplines into hard/soft; pure/applied; and life systems/ non-life systems, and Kolb (1981), whose much larger study placing academic fields on concrete/abstract and active/reflective dimensions resulted in a classification into four types of academic disciplines: social (e.g., education, social work, law); humanities and social sciences; sciences (especially engineering); natural sciences and mathematics. Biglan and Kolb both gathered their data through questionnaires: Becher (1986), in contrast, conducted a large number of unstructured interviews with academics across disciplines, and also arrived at four categories. hard pure (e.g., physics); soft pure (e.g., history, anthropology); hard

applied (e.g., mechanical engineeering); soft applied (e.g., education). Becher separates his categories on both epistemological and cultural grounds. This line of enquiry has, however, had little effect as yet on approaches to ESP. Strevens (e.g., 1977a; 1977b) has been one of the few to attempt an overview of ESP. The diagram below (Figure 3.2.1) is an attempt to define the spheres of operation of ESP: it combines Strevens' 1977(a) and 1977(b) diagrams and follows Strevens' suggestion that all ESP courses are either occupational or educational in nature, further dividing them according to when the course takes place:

Figure 3.2.1.



How specific is 'specific'? Strevens (1977b) describes four ways in which ESP courses may be specific:

> *The content of SP-LT courses are thereby determined, in some or all of the following ways: (i) restriction: only those "basic skills" (understanding speech, speaking, reading, writing) are included which are required by the learner's purposes; (ii) selection: only those items of vocabulary, patterns of grammar, functions of language, are included which are required by the learner's purposes; (iii) themes and topics: only those themes, topics, situations, universes of discourse, etc. are included*

> *which are required by the learner's purposes; (iv) **communicative needs:** only those communicative needs ...are included which are required for the learner's purposes. (p. 81)*

Clearly, a carefully-crafted ESP course for a well-defined context may be very specific indeed on the four categories above (e.g., Jones, 1978, Land, 1983). In regard to the EOP (English for Occupational Purposes) branch of the ESP diagram, courses may have to be very specific. There are good reasons why different jobs in the same country, or the same jobs in different countries, or similar jobs in different organisations, need their own English language course. There are also good reasons why many teachers are not equipped to do the necessary needs analysis, course design and materials development to deliver such courses themselves, and why commercial publishers are unwilling to invest in publications with such a restricted sales potential. Much ESP which may take place in 'educational' institutions is in fact occupational, but many teachers have begun to resist the over-specification of the course and attempt to include broader educational values and learning experiences (e.g., Carre, 1984).

The lack of a broad view is both a central characteristic and a central problem of ESP: an ESP solution is a local solution to the problem of providing English for the needs and purposes of a specific learner or group of learners, and this local solution can only be generalized to other learners with identical needs and purposes. Thus, ESP has frequently come to be seen as an ad-hoc, in-house approach, based on some sort of needs analysis, formal or intuitive, followed by a syllabus specification and frequently by the development of materials locally to meet the specified requirements. Because of this ad-hoc nature and diversified approach, ESP has been little informed by research. More recently there have been attempts to rationalize some of this diversity. Teaching ESP was in danger of becoming an art able to be practiced only by highly qualified native English speaker teachers: Robinson (op cit)

found in her survey that ESP courses are "normally undertaken by enthusiastic native speaker expatriates and not by local teachers".

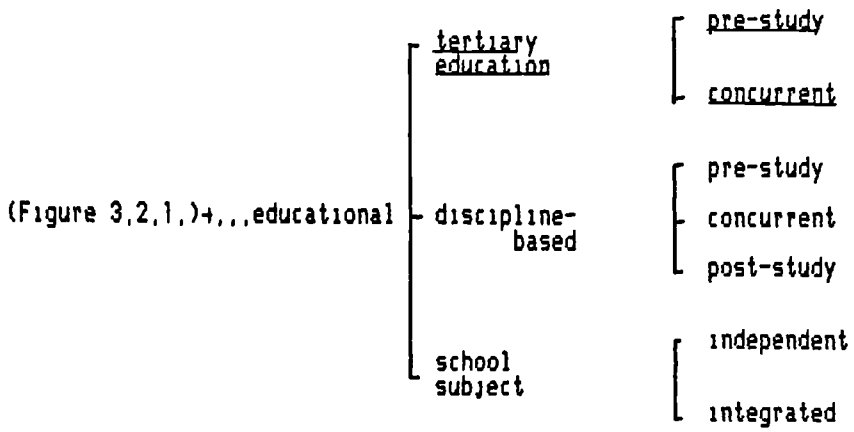
The tremendous strides which have been made in EAP (English for Academic Purposes) occurred in response to the evident lack of generalizability of most ESP, and the need to share some of the advantages of carefully-focussed teaching more widely than is possible when courses are very stringently ESP courses. EAP has developed as a significant branch of ESP only in the last eight years, as is clear when one looks back to Robinson's decription of EAP in 1980:

> *This is as yet an underdeveloped area in ESP, at least as far as published textbooks. ... Much work is going on in study skills and considerable research has been done (too often uncompleted, however). But most of this work, if accessible, is in the form of articles ... (p. 68-9)*

The EAP branch of the ESP diagram (Figure 3.2.1.) offers more potential for finding answers to the question of specificity which are simultaneously sufficiently specific to satisfy leaners' needs and sufficiently general to be logistically feasible in a wide range of contexts, and to satisfy some ofthe broader educational and social aims of teachers and learners. Such answers typically involve an adaptation of Strevens' division of EAP, which it is suggested here may be represented diagrammatically as shown in Figure 3.2.2. The addition (in bold) indicates the possibility (and increasingly, the actuality) of 'Study Skills' courses of a general and specific nature at tertiary level, that is, within universities, polytechnics, and further education colleges. A 'post-study' category is not included because it is suggested that such courses after tertiary study, if they were to occur, would be either discipline-based or within the EOP branch of the diagram (Figure 3.2.1.).

Such tertiary education 'study skills' courses are specific in skill terms (**restriction**) but general in content terms (**themes/topics**);

Figure 3,2,2,



they have some specificity of language patterns and rhetorical structures (selection), but few of vocabulary or language function; although they are specific to learners' communicative needs, these needs are themselves quite broad-ranging in non-content terms. Published materials such as Study Skills in English (Wallace, 1980), Panorama Williams (1982) and Strengthen Your Study Skills (Salimbene, 1985) offer a reasonable preparation for academic study, i.e. they propose a generalized academic purpose solution. Materials such as Reading Comprehension Course (Sim & Laufer-Dworkin, 1982), Study Listening (Lynch, 1983), Study Writing (Hamp-Lyons & Heasley, 1986) and Research Matters (Hamp-Lyons & Berry Courter, 1984) focus on a specific academic purpose defined in skills terms and propose that the skill is generalizable across academic disciplinary purposes. The earliest of these, by Wallace, was published only in 1980, which explains why Strevens' model did not, in 1977, take much account of EAP, and why Robinson could speak accurately of the lack of published materials in writing a book to be published in 1980.

The view which underlies these EAP materials posits that language proficiency is divisible into skills, and that certain language patterns and structures have a higher surrender value in study contexts than in other contexts, but that content which will be appropriate for all

learners, as language learners in academic settings, can be found. It posits also that while there are some specifics of lexis and language function which operate for every distinct SP, the development of a solid academic base of language control is prior and is more efficiently as well as more enjoyably done through courses focusing on the wide academic community within which language learners must live, than through narrowly focused, wholly instrumental courses. While such a view has less to commend it in EOP contexts, there are indications that language learners in such contexts also prefer a course which is not wholly instrumental (Mead, 1978).

2.2. Testing for specific and academic purposes

Although the term is quite new, specific purpose testing itself is not new, nor are its problems: the Schools Inquiry Commission was wrestling with the practicality and backwash problems of specific purpose testing in 1868:

> *When a school has to prepare boys for several different exams, an adaptation of the school course to suit them all becomes impossible. One boy, who is reading for the army, has to be taught one set of subjects; another, who is to be a medical student, another. It is easy, if the examinations are very stringent, to push this divergence between the different studies required so far, as to make effective organisation of the school, as a place of general education, impossible. (quoted in Brooks, 1984:p.324)*

We have seen in the preceding section that it is not possible to draw clear boundaries between ESP and EAP, but that it is a matter of degree of generalizability or specificity. The same is true for ESP/EAP testing. We may refer back to the distinctions between ESP and EAP made on the basis of Strevens' (1977b) model: the distinctions between ESP and EAP testing must be made on a continuum of specificity/generalizability of (1) restriction, (2) themes/topics, (3) selection, (4) communicative needs. But both ESP and EAP must be measured by the same fundamental criteria:

the expectations which language tests must fulfil, detailed in Chapter 1, must be fulfilled by ESP and EAP tests alike.

2.2.1. Reliability

We must remind ourselves that, to be valid, a test must first be reliable. ESP and EAP tests must satisfy the same reliability requirements as other tests. There appears to be a view that for ESP tests, especially when they are direct performance tests as in the case of oral interview and writing tests, reliability is not important, since these tests have inherent validity.

The PLAB (Professional and Linguistic Assessment Board, administered by the General Medical Council, and comprising language as well as medical sub-tests), for example, does not monitor reliability levels for the scoring of the writing test or the language interview, although it does require multiple marking for both. In Britain currently the ELTS (English Language Testing Service, administered by the British Council with the University of Cambridge Local Examinations Syndicate, and described in detail in Chapter 4 section 1) is the most widely administered and widely publicised ESP test. In the Specifications for an English Language Testing Service (produced by B.J. Carroll within the British Council in January, 1978, and reproduced in Alderson & Hughes, 1981) there is no reference to the reliability requirement for the ELTS. The ELTS does not monitor the reliability of its oral interview, and until recently did not monitor the reliability of the writing test: each of these direct tests is scored by a single rater. Skehan (1984), however, points out:

> *... the constructor of a language test is working in a clear tradition established in psychometrics, which accepts as fundamental that a test cannot be accepted as self-evidently good, but that its worth can only be established by examining its relationship with other performance criteria...*

Both the PLAB and the ELTS attempt a compromise over reliability by using a battery of tests, some of which are discrete point tests and subject to traditional item analysis: but the combination of a test score of high reliability with a test score of low reliability does nothing to raise the reliability of the unreliable score. In contrast, Test in English for Overseas Candidates (administered by the JMB, Joint Matriculation Board), which is an EAP test, includes a writing test and has each essay answer scored by four raters, with a fifth rater for borderline answers.

2.2.2. Validity

ESP tests have high face validity, as do EAP tests, although perhaps to a lesser extent, but validity of other kinds remains to be established. It is commonly claimed that ESP/EAP tests are more content valid than general tests. Clearly, the content of university courses varies as much in practice as is possible in theory, and the arguments for ESP tests are the same as the arguments for ESP syllabuses in this regard. Davies (1977) describes content validity as "an appeal to the subject expert" (p. 61). Criper and Davies (1986) point out that, whereas in assessing the content validity of general language tests there is only the language expert to be considered, in assessing the content validity of an ESP test there are two sets of experts involved: the language experts and the specialist subject matter experts (p. 111). Porter (1986) sees content validity as "concerned with the degree of fit to a theoretical model" (p.1), and it is here that there are problems with ESP/EAP tests, since there are not as yet any fully developed theoretical models upon which to base them.

The linguistic model underlying the ELTS, as stated above, is the Munby model, but, as Skehan (1984) points out, the model does not state the

relationships of the skills in the taxonomy to one another, nor their relative importance: thus the specifications for test categories are "no more than guesswork dressed up as a comprehensive theory" (p. 210). The specific purpose model underlying the ELTS, as put forward by Carroll (1981), was to take six hypothetical 'participants' representing overseas non-native English speakers, applicants for tertiary education courses in Britain, and to describe their needs of English on a number of 'specification parameters'. These descriptions were intended to guide the test constructors in selecting material for and in writing test items. It was from this beginning that the six separate strands or 'Modules' of the ELTS claimed their authenticity. Seaton (1983), however, tells us that "... what happened was that the specifications were edited down to a common core of tasks and skills; as if the six sets were placed on top of each other and someone looked down through them" (p.3). It can be seen from this that if there was any <u>specific</u> content validity in the specifications, it was washed out during the process of operationalising the test.

It would appear that there are real content validity problems with ESP tests, because the boundaries of each 'SP' have not been drawn, which is a logically prior stage to determining what the language content (E) should be. The TEEP (Test in English for Educational Purposes, Associated Examining Board) , which like the ELTS is a two-tier test, with a first component which is taken by every testee, offers only two choices in its second component: arts/social/administrative/business studies, and science/engineering. Porter (1985), speaking of TEEP, reports that testees want tests in their subject matter, or will tolerate tests with what they perceive as neutral subject matter, but that as soon as any hint of specificity is introduced, if this specificity is not the testee's own special field (as it is even more unlikely to be in the case of the TEEP than in the case of the ELTS) "anxiety levels rise and there is a feeling that justice will not be done." (p.3)

Anastasi (1982) describes the following procedure for establishing the content validity of a test:

> 1. *the behaviour domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions;*
>
> 2. *the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared;*
>
> 3. *content validity depends on the relevance of the individual's test responses to the behaviour area under consideration, rather than on the apparent relevance of item content. (p.132)*

In this view, content validity becomes both an *a priori* and an *a posteriori* activity, and the distinction between the degree of fit to a model (content validity) and the validity of the underlying constructs themselves (construct validity) becomes a fine one, as Weir (1986) points out.

Anastasi (op cit) describes construct validity as "... a comprehensive construct which includes all the other types." (p.153) Skehan (1984) considers the relevance of construct validity to ESP testing to be:

> *... it provides a link between theory and practice, which in the theory-practice direction, provides a more powerful explanation of testing procedure, and which, in the practice-theory direction, provides a way in which the underlying theory can be tested for adequacy. (p. 209)*

While, as we saw in the previous section, the evidence so far available seems to favour a partial divisibility hypothesis of language proficiency, we cannot conclude from this that the division along 'specific purposes' or 'academic purposes' lines is a true reflection of the way language proficiency divides, and from that argue a case for ESP tests. The case for ESP tests must be proved.

Skehan (op cit) criticizes the claims made for the construct validity of the ELTS on the grounds that the putative theory on which the test rests (the Munby model) is unsatisfactory internally and bears insufficient relationship with reality, and that the test has not been validated statistically. He contrasts the TEEP, for which careful validation procedures were followed in the test development stages. (p. 212) The externally-commissioned validation of the ELTS (Criper and Davies, op cit) found that on a principle components analysis a uni-factorial was most satisfactory for the ELTS: rotation of factors suggested a dominant first factor followed by a second (reading) factor and a third (listening) factor (p.130).

Henning (1986) asks why, if the ELTS is based on a multi-dimensional view of language proficiency, it is apparently a uni-dimensional test, on the validation study data? He asks whether it is a fault of the conceptualization of the test or of its construction, pointing out that when tests with low reliability are used in the correlation matrices for factor analysis, a clear factor picture is unlikely to emerge.

While the construct(s) underlying an ESP test should be statistically verifiable or falsifiable *a posteriori*, they should also have been fully and clearly worked out prior to test design and construction: otherwise, what is verified or falsified may not be the construct intended to be captured by the test. Reliability should also have been ensured in the development phase, so that in investigating validity the investigators are working with true variance rather than error variance. Even when the ESP test is adequately reliable, the problem with ESP testing seems to be that in general terms it is such an obviously right and sensible construct - but to go further and identify it precisely is an enormous, if not impossible task. By definition it is not a single construct but a complex of constructs.

Alderson (1981) points out that "*a priori* a specific test is impossible." (p. 123). It is not possible to construct a test for every describably different ESP, i.e., not only for architects, lawyers, dental technicians, etc., but also for architects whose focus of study is sociological and another for architects whose focus of study is aesthetic; and so on. We saw in sub-section 2.1. that there has been research into 'disciplinary cultures' and proposed categorisations of these: however, no single satisfactory categorisation has been reached, and certainly none of this work has yet been applied in ESP test design. There is not yet any evidence available to show why disciplines (SPs or 'subjects') should not be grouped together. On linguistic grounds the decision would need to made depending on the level of specificity at which the test was operating.

But the arguments for separate (ESP) tests relate more to the content than to the language, and in this regard, as Becher (1986) points out (p 4), there are opposing tendencies, one aiming to reduce the arena of investigation to a manageable size, and the other insisting on the recognition of important distinctions even within a single discipline. He tells us "To see the whole is to see it in breadth, but without access to the particular vision; to see the part is to see it in depth, but in the absence of the general overview." (p. 1) And this returns us to the group vs. individual dilemma of ESP/EAP testing again: at lower levels there is considerable overlap in content between different disciplines and therefore it would be possible to group students and administer them the same test. At this level, EAP tests, which typically treat all tertiary level courses together, dividing by skills but not by subject, may be most appropriate and most practical. This is the thinking behind the TEEP, which only makes a broad distinction between social sciences/humanities and science/technology.. At more advanced levels, however, Criper (1981) reminds us that it is necessary to account not only for the considerable variation from discipline to discipline but also for that among specialist areas of study within disciplines, if a test makes a serious

claim to be specific, and is aimed at postgraduate entrants. At this level, the more detailed aspects of selection, i.e., lexical and syntactic specificities, become important. Johns and Dudley-Evans (1980) found that specialist vocabulary was a key factor in students' failure to comprehend content area lectures. Houghton (1980), comments that:

> *... attention given to vocabulary represents something of a departure from current ESP orthodoxy, which ... tends to pay very little attention to vocabulary or to its grading for learning purposes. (p.26)*

When ESP first became a popular movement, as we saw in section 2.1., there was a move away from register studies and in particular from the teaching of content-area vocabulary. Attention was focused at the macro- or discourse level of disciplines. Recently awareness has grown that there are common patterns of discourse across disciplines, and there has been a shift toward genre analysis, or the linguistic and rhetorical analysis of discourse units. Swales (1986) suggests that for fairly large groups of disciplines, texts share "... regularized macro-structures and rhetorics that follow identifiable role-models" (p. 37) and suggests that it is at the lexical and syntactic levels that disciplines differ most. This view, applied to ESP tests, would suggest that the use of highly specialised lexis and a high frequency of certain syntactic structures would provide the greatest 'authenticity', i.e., face validity, and also the greatest content validity. It would also provide the greatest problems for any students incorrectly assigned to the particular branch of ESP of that test.

Alderson and Urquhart (1985) studied the influence of students' academic discipline on their performance on a range of ESP reading tests, they concluded that, although academic disciplines can play an important role in test performance, the results were not consistent and the inconsistencies could not easily be explained. They also felt they had found a need to take other factors, notably linguistic proficiency level and test item type, into account. Their results might be explained by

marked effects of specific lexical items and syntactic patterns on individual students, an explanation which would accord with Criper's comments, above, and with the findings of Johns and Dudley-Evans (op cit) and Swales (op cit).

If it is true that it is these lower-level features which distinguish disciplines from one another, any test which purported to be discipline-specific, i.e., an ESP test, would need to show that it distinguished in this way, and did so correctly, before it could lay claim to construct validity. But it would also need to be demonstrated that these differences are meaningful: that is, that students' test scores really are affected by the selection of different specialist lexis and syntax. Research in the social construction of knowledge (e.g., Knorr-Cetina, 1981) suggests that disciplinary communities are formed in ways which have little basis in linguistic similarities or differences Alderson and Urquhart were unable to account for their results, but they took academic discipline as a given and did not draw their own distinctions: indeed, we are not yet in a position to do so, and must await the results of research in other fields, notably in the epistemiology and sociology of knowledge and in genre analysis.

A great deal of work still remains to be done in the construct validation of ESP tests. Because of the formidable size of the task, an approach parallel to the one currently being taken in genre analysis seems a fruitful one: that is, to focus on one aspect of the construct of one ESP test, and investigate it fully. Ideally, this would be done through both *a priori* and *a posteriori* methods, ensuring that any statistical procedures applied would be confirmatory rather than exploratory. If any support is found for the construct in that instance, it would provide the impetus for other studies of other aspects of the same and other ESP tests. Given sufficient studies and replications, such a 'triangulation' approach could eventually lead to a fuller understanding of the construct underlying ESP tests, and lead to better testing practices.

2.2.3. Achievement, proficiency or diagnostic testing?

In theory, ESP achievement tests present few problems: instead of assessing whether learners had mastered 'the language' it would be possible to assess whether they had 'satisfied their needs'. Instead of:

(a) the language → the syllabus → teaching → student assessment

we would have:

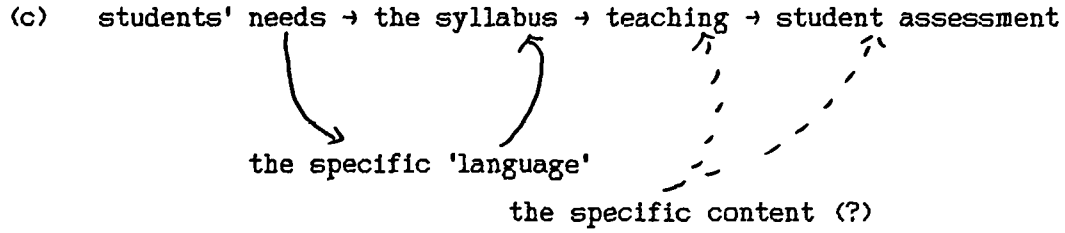
(b) the language → } the syllabus → teaching → student assessment
students' needs → }

The pattern applies to EAP achievement testing, with the difference that students' needs are less precisely defined and thus less precisely assessed.

However, while ESP achievement testing is appropriate in-house, when the context of testing is one where students come from a wide range of backgrounds and must be assessed for their suitability for placement on a wide range of courses, the concern must be with ESP proficiency testing, that is, with testing without knowing the syllabus the learner has followed. This context, a proficiency testing context, is the one within which the ELTS, TEEP, JMB and PLAB all operate, though at different levels of ESPness.

ESP proficiency testing presents many more problems: not only do the same sampling problems occur, but it is not possible to specify the universe which must be sampled as it is with ESP achievement testing. There is no such thing as, cannot by definition be such a thing as, 'absolute proficiency in ESP'. ESP proficiency must always be relative to something: but that 'something' is not a syllabus. Rather, it is a

'language', the specific language of the specific purpose or use. The model is not so simple:



The model shows that ESP 'proficiency' tests may be better used for diagnostic purposes than as true proficiency tests. Because ESP tests, like achievement tests, are always tied to a known (or at least, knowable) quantity, the information from them may be used to find out not simply to what extent the student has already mastered what he needs for the specific purpose, but also which of the needs the student has not yet mastered. Alderson (1981) says:

> *Perhaps the most powerful argument for specific tests is that of the diagnostic value of a profile of a student which can be matched against the communicative needs of his particular course of study. ... there ... may be value in profiles of students' abilities, relatable to institutional criteria, for both administrative purposes (that is, for admissions decisions) and for pedagogic purposes, since hopefully such information would allow remedial action to be taken on a language course, for example. (p. 125)*

The use of the information from a test to make individual decisions makes the test more 'specific', i.e., brings it closer to the elusive ESP-ness we have been seeking in this section. The implication of the use of detailed information from the test is that the information must be interpretable by score consumers, who are primarily admissions officers and academic faculty. Criterion-referenced scores, as we saw in Chapter 1, section 2, are more easily interpreted than norm-referenced scores, and this is one argument in favour of their use with ESP tests. Further, if there are important differences from discipline to discipline, and between groups

in the 'same' discipline; and if, as Alderson (op cit) suggests, it will always be impossible to construct ESP tests for each of those different contexts, then the solution would ultimately seem to be to provide score consumers with very detailed information about test items and scoring criteria, and permit them to set their own criterion for test results as a whole.

The discussion has not been exhaustive of the potential types of ESP and EAP tests: it is possible to have 'wide spectrum' proficiency tests which straddle the ESP/EAP divide, and the TEEP does this; EAP proficiency tests are also possible, and the JMB is an example. When the tests also assess each skill area separately, as the ELTS, the TEEP and the JMB do, there is specificity on the skill dimension - but this is not what is usually meant by a 'specific' test.

2.2.4. Practicality

Given the large claims that have been made for ESP teaching, and the amount of money which has been spent on ESP course design and materials development around the world, it may seem surprising that there has been so little attention paid to ESP testing. The explanation lies at least partly in their apparent impracticality. Sinclair (1979) justifies the lack of attention to any aspects of testing and evaluation on practicality grounds:

> *A fully reliable test instrument is a major project in itself requiring specialised staff. It would only be worth the effort if it was usable well beyond the present circumstances, and this would be of a rather general character. This requirement runs contrary to the often specialised nature of ESP, requiring a compromise in design. (p.114)*

Skehan (1984) reiterates Sinclair and adds more powerful arguments when he says:

> *The issue of producing tests for small groups of people is particularly relevant to the issue of assessment since one cannot, practically speaking, justify the production of a test for one individual. Tests are time-consuming to prepare, require piloting and revision, etc., and therefore have to be given to a group of people (and probably to several groups at different times) if they are to justify the preparation and analysis that is necessary to produce a test reliable and valid enough to be the basis for important decisions. (p. 206)*

It cannot be denied that when there are many tests instead of one test the demands of time and money are greater. ESP/EAP test construction also demands specialist constructors, and involves a range of experts, probably different ones for each 'track' of the test battery. Such constraints were not accepted by Carroll (1978) however, in the design of the ELTS:

> *... we will bring to bear on the test design important operational considerations affecting the administration of the test service, but it must be emphasised that such considerations, however pressing, will not make the communicative needs of the participants disappear. We would hardly be likely to achieve our aim of test improvement if we ignored a patently essential communicative need merely because it entailed practical decisions. (p. 67)*

It is necessary to set against the cost of an ESP test the various uses to which it can be put. Cronbach (1971) points out that every time the scores on a test are used, its utility value increases: thus if Alderson's suggestion (above) is accepted, the scores on a test such as the ELTS may be used at least twice (once for making placement decisions and once for diagnosis and planning of remediation where necessary), thus doubling the utility of the test. If they are used in sophisticated ways as suggested above, very exact placementdecisions can be made, and the utility value is further increased. However, it seems likely that ESP tests cost more than twice as much as traditional standardised general English tests to develop and administer, especially given the inclusion of direct tests in ESP tests. Every component in the ESP battery is

effectively a separate test for development purposes, and the number of sub-tests decides development costs, administrator time, scoring costs, and complexity of score reporting. It is not strictly appropriate to compare the practicality of ESP tests with standardised tests, however, because in the current climate in language testing the alternative to an ESP test would more likely be an EAP test, which would include direct performance tests just as the ELTS, PLAB and the TEEP do. As comparative data is not available on the research and development nor the administrative costs of any of these tests this discussion must remain hypothetical.

2.2.5. Backwash

As we saw in Chapter 1, the term 'backwash' refers to the effects of a test on curriculum. Backwash can only be judged as positive or negative from a relative standpoint, that is, from an existing set of values, and if there is a prior commitment to ESP teaching then presumably the backwash effects of ESP testing will be judged beneficial, because an expansion of ESP teaching can be expected to occur. We saw above, however, that there have been reservations expressed about too great a concentration on the ESP in ESP courses to the exclusion of the socio-cultural aspects of the language (further examples are Jordan and Matthews (1980) and Chamberlain and Flanaghan (1980).

The backwash from ESP achievement tests is likely to be less flexibility to include more general topics and interactions in the classroom and the increased centrality of ESP materials in classroom discourse, thus reinforcing the "peculiarities" (Phillips and Shettlesworth, 1980) of the ESP classroom. Swales (1984) says:

> *Students ... look to the ESP classroom for certain personal values, wherein their role as real people with real interests can get greater recognition than elsewhere in the learning environment. (p. 14)*

However, the backwash from an ESP proficiency test such as the ELTS or TEEP, as entry tests rather than exit tests, may well be a sharpening of focus and careful attention to objectives in general English classrooms. When the diagnostic potential of these tests is used it is also likely to have the effect of increasing the number of carefully-planned courses designed to meet the language needs of students who have been diagnosed as lacking in specific skills. The backwash from EAP tests is likely to be increased teaching of or greater attention to 'study skills' on general courses, and the provision of more courses designed especially to prepare tertiary education applicants for life in the university. All of these seem to be desirable effects, but this assumes that all the learners who are caught up in the 'backwash effect' do in fact have these desires and needs for English. Here we return to the group vs. individual dilemma which was discussed in the construct validity section. not every learner is necessarily in a position to know his desires and needs, still less to be able to negotiate them with teachers and testers.

3. RELATING 'GLP' TO ACADEMIC/SPECIFIC ACADEMIC PURPOSE TESTING

We have seen in this Chapter two approaches to the same problem: how is language proficiency portioned out within and among individuals, and what difference does it make? We may characterize the argument about general language proficiency as theory-driven, while the arguments for ESP are primarily practice-driven. In other words, the GLP approach has been to say 'here is a problem: can we find a solution?', and then to conduct extensive empirical trials to test out a range of possible solutions. In contrast, the ESP approach has been to say 'there is a problem, of providing adequate language instruction for the needs of particular groups of learners, and of measuring their English proficiency: here is a solution', and then to conduct armchair-research into a pre-conceived solution. Skehan (1984) highlights the contrast in approaches between these two:

> *The two approaches, that characterizing research into the unitary competence hypothesis, and that relating to ESP test development, have produced widely different outcomes The former has used traditional psychometric methods and has made slow but steady progress; the latter has put its faith in a new approach to applied linguistics, or rather one particular exemplification of this, and has encountered serious difficulties. The unitary competence hypothesis has had to face some challenging evidence; while it has had to be modified considerably, Oller can claim that stating the hypothesis has led to a considerable quantity of research which has extended our knowledge of language proficiency. Proponents of the ELTS, in contrast, have difficulty in demonstrating the empirical basis for the new form of the test, with the result that we are left with very little progress indeed of any solid nature - a few testing techniques, perhaps, but no information about whether the test meets the claims that are made for it, and where future research should be directed. ... the 'surrender value' of the two approaches has differed markedly. (p. 213-4)*

The ELTS and ESP testing are not synonymous, of course, but a further problem with the ESP approach has been that there have been few people brave enough, or with sufficient resources, to undertake ESP testing other than on an ad-hoc, in-house basis. The judgement ultimately passed on the ELTS is likely to be the principal factor in the judgement passed on ESP testing as a whole (Henning, personal communication).

It is now possible to turn to the context of this study, applying what we have learned to a consideration of tests of writing, attempting to understand what would characterize an ESP writing test as distinct from an EAP writing test.

4. WRITING FOR ACADEMIC PURPOSES AND SPECIFIC ACADEMIC PURPOSES AND ITS TESTING

4.1. Writing in academic settings

There is currently considerable interest in the notion of 'discourse communities', a notion which brings together language, language user and language use in dynamic ways, particularly the study of writing as it is used in the university community and in specific disciplines (e.g., Bazerman, 1981; Myers, 1985; Herrington, 1985), as a means both of understanding the discipline and of understanding the writing processes. Herrington (1986) used analysis of student texts, interviews, and classroom observation to try to understand how students learn to see themselves and to function as members of academic discourse communities. Although Herrington's research indicated differences in knowledge structures represented in writing in different disciplines, she also believes she has been able to identify some characteristics of good apprentice writers which are common across disciplines: successful students perceived that in their writing they needed to create an issue for themselves and work to resolve it, first, for themselves and then, to convince their professors; they saw themselves as an audience, in the sense of using their writing to explore an issue and shape their responses, and in the sense of convincing themselves they had resolved the issue to their own satisfaction; they were able to interpret and act upon the information they received from their professors in getting closer to an understanding of the disciplinary culture, even though much of this was vague, professors leaving tacit the most important values of the discipline.

The work in the previous paragraph was with native English-speaking students: we may expect initiation into the university community in general and into specific disciplines to be much more difficult and

tentative for students who are attempting to master the language at the same time. Ballard and Clanchy (1986) describe the growth towards literacy in academic settings as "cracking the code, mastering the alphabet of linguistic and cognitive behaviour", where "...A, for example, introduces key elements in the academic culture, as well as standing for Excellent at the end of an essay. A, the student learns, stands for *Analysis*... also for *Argument* (and) *Assertion*..." (p. 6) Ballard and Clanchy discuss some of the elements of literacy in the academic culture and show how academic faculty, through the judgements they make on students' writing, seek explicitly or implitly to acculturate them.

Ballard and Clanchy, and Herrington (1986), both point out that for undergraduate students, growth through the university years involves attaining control in more than one discipline at the same time; although this may be more true in Australia and the U.S.A, than in Britain, Becher (1986) found that even in Britain real initiation into a single discipline normally occurs only at postgraduate level. Weir (1983), conducting the research for the TEEP, investigated the content of university postgraduate courses, and found tremendous variety, not only from university to university or from Faculty to Faculty, but even within departments, in terms of the types of tasks required of students, the expectations for student performance on tasks (both level and criteria), and the allowances made (or not made) for non-English native speaker students.

Phillips et al (1985) conducted a detailed observational study of faculty adjustment to Indonesian students entering tertiary education in Australia, and found tremendous individual variation among faculty in the way they responded to students, the amount of time they spent with students academically and socially, and the amount of leeway, if any, they gave them because they were non-native speakers of English. Faculty apparently vary in their perceptions of students as members, or potential members of their own discourse community, and either smooth or bar their entry.

The recent work in genre analysis which cuts across disciplines to look at, for example, article introductions (Cooper, 1981; Swales, 1981 & 1984(a); Dudley-Evans, 1983 & 1986), can be seen as a search for features of discourse communities which are generalisable across disciplines, and therefore teachable in EAP or 'broad spectrum' ESP courses rather than in a multitude of independent 'narrow-spectrum' ESP courses. A focus on common patterns of text structure and internal logic minimises both the difficulty the learner may have with the language of the subject content, and the difficulty the language teacher, as a non-member of the discourse community, may have with the subject content of the language. Discussing research paper introductions, Swales (n.d.) suggests that:

> *...appreciation of the schematic, purposive and developmental expectations... has, in alliance with student expertise in their individual content areas created a set of texts to which we can ascribe a sufficient amount of coherence to overcome uncertainties that may otherwise have arisen as a result of developmental or registral mismanagement. (p.3)*

The genre analysis approach can be seen as a move from a focus on ESP toward EAP, i.e., from specifity to generalizability, but through an informed, research-based, understanding of what it is that unites discourse communities and can reasonably be taught by language specialists, and when it is necessary to stand back and let the students who are members, or apprentices, of that discourse community depend on their own expertise.

4.2. Testing writing in academic settings

We have seen in this Chapter that the validity of the divisible skills view has been extensively researched in recent years and has been at least tentatively vindicated. In contrast, the assumptions behind English for Specific Purpose have yet to be satisfactorily researched. In considering the testing of writing in academic settings, the question asked earlier must be asked again: how specific is specific?

We will be concerned in the ensuing chapters with the attempt to investigate the construct validity of a writing test which is claimed to be related to the needs of different disciplines. In order to evaluate that claim we need to understand, as far as this is possible in the current state of knowledge about this construct, which has been explored in section 2, the difference between a writing test which is discipline-specific and a writing test which, although academic, is discipline-free.

4.2.1. Levels and assumptions

Both types of test which will be investigated are tests of writing in academic settings, but they differ in terms of the level of specificity they attribute to the concept 'academic setting', and in terms of the assumptions they make.

As we saw in sub-section 2.1, 'academic writing' is a tertiary education study skill: in Figure 3.2.2. it may be placed in either the pre-study or concurrent sub-categories. It is viewed as specific in skill terms but as general in academic discipline terms. In the ensuing chapters a test of writing for general academic purposes is referred to as a 'GAP' writing test. A GAP writing test is based on the assumption that language proficiency is divisible along a skills dimension, i e., that a testee cannot be assumed to have a certain level of writing skill because she or he has that level of skill in reading, and/or speaking, and/or listening. It is based on a further assumption, that proficiency is general in discipline terms, i.e., that a testee will perform as well on a writing test targeted to the university community as a whole, as on a writing test from her or his own field of study.

As we also saw in sub-section 2.1, 'specific academic writing' falls into the 'discipline-based' category (Figure 3.2.1), whether pre-, in- or post-study. It is viewed as differing from 'academic writing' in the level of specificity attributed to the academic discipline dimension. In the

ensuing chapters a test of writing for specific academic purposes is referred to as a 'SAP' writing test. A SAP writing test is not only based on the assumption that any testee may have different levels of proficiency for the different language skills, but on the additional assumption that any testee may exhibit different levels of proficiency on a writing test targeted to the university community as a whole and a writing test from her or his own field of study. Within the ESP construct there is a further assumption, that it will be easier for the testee to perform well on a SAP writing test than on a GAP writing test.

4.2.2. Implications

The investigation in the ensuing chapters is designed to investigate these assumptions. If it is shown that SAP writing tests do indeed meet the claims made for them, there will be important implications for future design of writing tests for academic settings. Resources will need to be assigned to the development of a large number of separate writing test instruments to match each distinct disciplinary need. A prerequisite of this will be extensive and intensive research of the writing of disciplinary cultures in order to determine at what point two 'subjects' diverge and a new SAP writing test becomes necessary. Task design will become critical, and may not be able to be done by language specialists without considerable use of subject specialist informants. Test administration and scoring will take on new dimensions as testees select from an ever-increasing menu of specific choices. Serious thought will need to be given to the testee's right to select from the menu, even if this means she selects inappropriately. Raters will need to be trained to score each test reliably and validly, and consideration will need to be given to whether language specialists are the best people to be scoring SAP writing tests. Decisions will need to be made about the ways in which the scoring criteria and procedure reflect the specificity of the test.

CHAPTER FOUR

SPECIFIC ACADEMIC PURPOSE and GENERAL ACADEMIC PURPOSE WRITING TESTS: AN EMPIRICAL INVESTIGATION

1. BACKGROUND TO THE STUDY

1.1. Context

The study reported in this Chapter is set in the context of a larger-scale study of the English Language Testing Service, the testing service provided by the British Council. The British Council is Britain's principal funding body for scholarships for overseas applicants to education institutions and courses, and the English Language Testing Service (ELTS) as originally introduced and operated was a test for applicants to academic tertiary educations (universities, polytechnics, and advanced professional courses). Because the undergraduate entrance requirements of British universities are difficult for students from other educational systems to fulfil, almost all applicants are postgraduates or equivalent.

The ELTS is a two-tier test. In the first tier there are two multiple-choice tests, one of which is a reading test and the other of which is a listening test. These two tests are taken by every ELTS testee. In the second tier there are three tests: a multiple-choice study skills test; a direct writing test; and an oral interview. In this second tier there are six choices or 'Modules', and every testee must choose one of these. The six Modules are General Academic, Life Sciences, Medicine, Physical Sciences, Social Studies and Technology. The last five of these were designed to conform to the five largest groups of applicants for British Council scholarships, while the first, General Academic, was designed for all those applicants who did not fit easily into the other Modules. The

test went into operation in 1980 with approximately 4,000 candidates and in 1985 had some 10,000 candidates.

The British Council provides professional oversight of the test in terms of expertise in testing English as a foreign language, English for Academic Purposes and English for Specific Purposes. Their partners the University of Cambridge Local Examinations Syndicate (UCLES) provide technical expertise with multiple choice item writing, print and distribute test materials, and store all score data and student records in their computer database. Test administrators are provided and overseen by the British Council; a different section of the British Council receives and acts upon scores before these are sent to UCLES.

The ELTS was introduced to replace the English Proficiency Test Battery (EPTB), which had been used since the mid 1960s, for a number of reasons. First, as a short, multiple-choice test, only four versions of which existed, which were continously re-used, the security of EPTB was always at risk and was being seriously questioned towards the end of the 1970s. Second, the 1970s had brought a new paradigm in language teaching, a humanistic and communicative paradigm in which student needs and differences were being stressed. This paradigm was also influencing language testing, and the British Council wished to both take advantage of the new insights and put itself at the forefront of the field. One of the key developments in the same period was the field of English for Specific Purposes (ESP), which as was shown in the previous chapter is closely related to the same general paradigm. The new testing service was intended to be a more finely-grained instrument which, according to Carroll (1978), would provide two kinds of information about any applicant for tertiary study in Britain:

> *whether he is already likely to be able to meet the communicative demands of a given course of study, or, alternatively, what would be the nature and duration of the course of language tuition he would need in order to reach the required competence level. (p.4)*

The ELTS was designed to put into practice three theoretical positions or constructs of how language proficiency is composed. The first is relatively uncontentious: the ELTS views language proficiency as divisible on the skills dimension, and it has separate tests of reading, listening, writing, and speaking. The ELTS takes this construct further than other tests, certainly further than any other test at the time of its introduction. It divides each of the 'objective' tests into items which test specific skills and 'micro-skills', and provides specifications of which skills or functions each item is testing (e.g., Skill 21: expressing information explicitly; Skill 25: expressing conceptual meaning, especially (e.g.,) micro-skill 25.1: quantity and amount) or 'micro-functions' (e.g., Function 5: argument; sub-function 5.1: information; micro-function 5.1.1.: stated/asserted (state, inform, tell, express, report, etc., etc.). These specifications are based on the needs analysis work of John Munby, who completed a Ph.D. under the auspices of the British Council at the time the British Council was considering the introduction and possible design of a new test. The central part of his study is a 'communicative needs processor', in which all the 'micro-skills' and 'micro-functions' are listed and grouped (the examples above are taken from this): a major part of his work was published as Communicative Syllabus Design, (1978)

The second construct underlying the ELTS is a view of language proficiency as divisible into 'general' and 'study' proficiency. The first, general section of the test "tests listening and reading skills and is intended to test general ability in the use of English"; the second tier is a Modular section which "tests language study skills used in reading, writing, listening and speaking and is related to a specific subject area" (ELTS: An Introduction, British Council/University of Cambridge Local Examinations Syndicate, n.d.).

Third, and perhaps most contentiously, the ELTS divides language proficiency on a subject, or discipline, dimension, through the six Modules referred to above. We may think of the three sub-tests for each Module within the second tier of the test as forming an ESP test battery. The justification of this kind of division by discipline is "... the hypothesis that the solution to our testing problem ... is through the diversification of test instruments to meet the diversity of the test situation." (Carroll, 1978: p.4).

Little has been published to aid an understanding of the design of the ELTS. The only publication which goes into any detail is Brendan Carroll's Specifications for an English Language Testing Service, prepared for the British Council's English Language Division, presented to that body in January 1978, and reprinted in Alderson and Hughes (1981). In that document Carroll applies Munby's taxonomy of skills and functions to the construction of 'profiles' of six participants in courses of study in British tertiary education institutions. These participants are not real people, but hypothetical people constructed by a number of "compilers", by reference to "contacts" and "documents" (p.8); the courses similarly are not actual courses but constructed ones. In Carroll's words, "we decided that less time-consuming methods would be sufficient to assess the basic adequacy of our approach to test specification" (p.7).

Carroll's application of Munby's taxonomy to the 'data' on his six hypothetical participants revealed a network of relationships among his proposed 'courses of study' which led him to the conclusion "that Language Skill requirement patterns cut right across disciplinary boundaries" (underlining in original) and that "the smallest communicative relationships (occur) between disciplines which seem to have the most in common" (p.9). This conclusion appears to contradict the argument upon which Carroll built his claims for the need for a test like the ELTS. Carroll is apparently aware of this difficulty:

> *Even if the... programmes are highly correlated communicatively, it still remains that the spoken and written discourse of the disciplines are very different indeed; their linguistic and diagrammatic realisations have very different appearances. (p.19)*

As we saw in the concluding section of the previous chapter, the evidence is not available even now, eight years after Carroll wrote the Specifications, to support this claim. The problem here appears to be that Carroll bases his network of relationships solely on the outcome of the application of the taxonomy. Information is not available as to how Carroll set about applying the taxonomy, but Criper and Davies (1986) show that application of the Munby taxonomy to the ELTS objective items as these were eventually written poses many problems (p.102-113). Most importantly, there is no perfect match between actual items and Munby's categories: some items require multiple categorisation and others defy categorisation by the taxonomy. Further, some categories in the taxonomy cannot be distinguished from one another, and the level of precision varies tremendously from item to item. Even when items can be categorised by the Munby system it cannot be actually known that in succeeding or failing on a particular item the testee was in fact using the language behaviour specificied for the item: such a system cannot account for individual learning and problem-solving strategies.

Carroll does not produce a network of events/activities based on his application of that part of the Munby model: had he done so (and had his data been actual rather than hypothetical) a different pattern of relationships might have been found, one less counter-intuitive. As it is, he attempts to justify the continued pursuit of discipline-specific tests:

> *Can we then test different disciplines with identical test material, selected to test their common communicative requirements? Or will we, in doing so, use over-generalised language/diagram realisations which may favour candidates in one particular discipline or, worse still, be equally irrelevant to all the disciplines? We are not yet in a position to answer these questions, so we propose*

> *to continue in a pragmatic fashion by preparing tests in different disciplinary areas and by paying particular attention in test data analysis to assessing any benefits, in improved test effectiveness, which can be related to diversification on a disciplinary basis. (p.19)*

Clearly, the introduction of such an innovative and potentially influential test by a body with so much real-world decision-making power necessitated an externally-directed validation study by leading experts in the field. Summarising the discussion of the ELTS at a meeting of such experts immediately prior to the operational introduction of the test, Alderson (1981a) said:

> *... it is crucially important to find out what is happening on a test as influential as the ELTS test. There is a clear need to know how such 'ESP' tests relate to existing tests, for academic as well as practical reasons. There is clear need to know what sort of diagnostic information can validly be provided, and whether it can be used by both applied linguists and lay people. (p. 133)*

The Edinburgh ELTS Validation Project was commissioned and commenced in 1982 under the direction of Alan Davies and Clive Criper. The Edinburgh ELTS Validation Project (EEVP) was to run from September 1981 to March 1986, and was extended until August 1986. The Final Report of the EEVP was presented to the sponsors, the British Council and UCLES, in September 1986 (Criper and Davies, 1986). This researcher was half-time Research Associate with the Project from October 1982 to September 1984 and from October 1985 to March 1986.

A test such as the ELTS, and a validation study such as the EEVP, offer fruitful ground for the pursuit of individual research: for this researcher, however, a long-term interest in the teaching and assessment of writing in both English as a first language and English as a foreign/second language meant that M2, the writing test, was the obvious

choice. The balance has swung so fast and so sharply in favour of direct tests of writing over indirect measures of writing-related ability, that poor reliability is often accepted on the grounds that it is compensated for by good validity. There has been a tendency to assume an inherent validity for direct tests of writing: they test the 'real thing', after all. However, when the issue is the choice between 'academic' writing tests of varying levels of specificity and 'general' writing tests, the assumption that the test is valid must be questioned and tested.

1.2. Design ofthe ELTS writing test

Close study of the 1978 Specifications reveals little to suggest how M2 was designed and constructed. It is clear, however, that the basis of test design was intended to be the close linking of the testing service with the "communicative demands study programmes make on the participants" (p.6). To arrive at specifications the compilers (members of the staff of the British Council's English Language Division) contacted institutions or individual specialists in the disciplines of the six hypothetical participants. no detailed information is available as to the kinds and quality of advice received nor what account was taken of it The Specifications state that "continual reference was made to authentic documents in the disciplines such as College Handbooks, Course Syllabusues and standard subject textbooks", but there is no statement that the texts included in the Source Booklets, the texts on which the questions for the writing tests are based, are themselves authentic.

Examination of the profiles of the six participants (Appendix A, Specification 7: Events/Activities) reveals widely differing amounts and types of writing needs from one participant to another, but these differences are not discussed in the body of the document. The intention of the test design was that testees would be matched with courses of study, and that if, for example, a course of study called for little writing, the testee's score on the writing test would play little part in

decision-making. This intention has not been put into practice to date, but even if it had been, it would remain the case that the finding of widely differing types of writing needs in different courses of study should result in widely differing writing tasks on M2 if a claim to validity is to be upheld. However, the Specifications do not recommend differing M2 tasks, saying only this in regard to the design of M2:

> *Writing Skills test; problem-solving, descriptive and reference skill writing based on information booklet. (Subjective rating according to scale and with photo'ed samples of examples at different levels.) (p.25)*

The actual tasks on M2 cannot be matched with Carroll's six participants and their needs because the eventual configuration of Modules is not the same as Carroll's six categories. An understanding of the design of M2 must be based on study of the actual test items, since there is no document available which discusses the design of the writing component Investigation of the M2 tasks suggests that these are neither deliberately the same as nor deliberately different from each other. What is consciously the same is that in every case M2 consists of two tasks, one of which should be completed in 25 minutes and one which should be completed in 15 minutes. The second question requires limited selection and transfer of information from an input text, and has come to be known as 'convergent'. In contrast, the first question is referred to as 'divergent', because although it is linked to an input text testees are asked to bring in something from their own knowledge or experience, or to give an opinion.

The implicit claim for M2 is the same as that upon which the ELTS as a whole is based: that the division into a number of distinct test tasks on a disciplinary dimension will enable a better matching of testee to test and will result in improved information (Carroll, 1978: p.4).

1.3. Definition of terms

In the study which follows, two terms to distinguish two kinds of writing tests are frequently used and must be defined. A 'general academic purpose' writing test is defined in parallel with 'academic writing' as that term was used in Chapter 3, Section 4.1. A 'general academic purpose' writing test by this definition tests writing as a general study skill, and the assumption underlying 'general academic purpose' writing tests is that a testee will peform equally well or badly regardless of whether the topic(s) of the writing test address content from her or his own field of study. Throughout the discussion which follows 'general academic purpose' writing tests are referred to by the acronym 'GAP'

A 'specific academic purpose' writing test, in contrast, is defined, also in the terms described in Chapter 3, Section 4.1, as specific to different academic disciplines. The assumption underlying 'specific academic purpose' writing tests is that a testee will perform differently on a writing test which calls on the testee's knowledge of her or his own field of study than on a GAP writing test, and the implication is that the difference will be in the direction of a more favourable test score for the testee. Research reported in Chapter 3, Section 3, indicates the difficulty of accurately delineating the boundaries of the specific academic disciplines. The investigation reported here uses the parameters of areas of study established for the British Council's ELTS. Because the object of study is the writing sub-test of that test, the parameters used for that test are also part of the object of study. In the model of a specific academic purpose test set up by the British Council for the ELTS, and applied in the design of the Modular sub-tests, there are five clearly different 'specific academic purposes': Life Sciences, Medicine, Physical Sciences, Social Studies, and Technology; plus a 'General Academic' division for "candidates whose area of study is not covered by one of the other Modules" (ELTS: An Introduction, British Council, n.d.). The research basis for this delineation has not been

reported and is therefore open to question. Throughout the discussion which follows 'specific academic purpose' writing tests are referred to by the acronym 'SAP'.

1.4. Objections to ELTS M2

When the ELTS was being developed, interest in the direct testing of writing was growing in general, and the British Council's commitment to ESP testing and to communicative testing meant that writing really had to be tested through direct performance. At the same time, however, language testing was undergoing something of a reaction to the psychometric-structuralist period (as discussed in Chapter 1) and reliability was lower on the agendas of many test developers than validity (in the restricted, non-statistical sense of conforming to a certain view of how the test should be). In the case of M2, the concern of the test developers was primarily with content validity, and questions of score reliability were little considered. Rather, it was assumed that because the procedure used to assign scores on M2 was 'criterion-referenced', reliability was not an issue: Seaton (1980) said:

> *In the very difficult area of language skills covered by M2 ... it was essential to devise an entirely reliable system of marking. It was at this point that the advantage of criterion-referenced testing proves its worth most effectively. M2 ... could be marked by consulting the scale of student performance ... (p. 112)*

However, while ELTS M2 is a direct performance test, it does not necessarily follow that the scoring procedure is criterion-referenced within the definition of that term as discussed in Chapter 1, Section 2.3. Within two years of the introduction of the ELTS, objections were being raised to the writing test based on anecdotal evidence and impressions of its poor reliability.

Objections to writing tests on reliability grounds are by no means new, and for some of us their recursiveness adds to rather than detracts from their interest. A much newer objection, though, was to the introduction of the ESP construct into the test. For the writing test, objections on this ground may be described as based on two arguments. the first is, "there's no such thing as discipline-specific writing - all academic writing is pretty much the same". If this objection were true, it would follow that discipline-specific writing tests are not necessary, since they would provide no additional information over general academic writing tests. The second is "of course writing is discipline-specific, but the writing tests in the Modules do not represent that specificity". if this objection were true, M2 would be invalid and scores could not be depended upon, however reliable they might be.

The second argument divides into two sub-arguments: the possible totality of disciplines is incorrectly parcelled out, or, the writing required in the test is not sufficiently like the writing required in the disciplines. If the first of these sub-arguments were found to be true, there would be important implications for the design of the ELTS; the Modular breakdown would need to be altered - and similar studies would need to be carried out for the other two Modular sub-tests since it could be predicted that they would equally be found to be incorrectly parcelled out. If the second of the sub-arguments were found to be true, greater attention would need to be paid to establishing a 'fit' between the writing tasks in disciplines and the writing tasks on the test; the issue of the level of specificity would be brought to the forefront.

In the investigation which is reported in this chapter, two of these objections are treated as problems to be studied. First, an operational problem. were ELTS M2 scores as used operationally sufficiently reliable for ethical decisions to be made about testees, i.e. about scholarship applicants; were they sufficiently for use in the investigation of the test's validity? Second, a theoretical problem: was there any evidence to

support the view that testees would write qualitatively different essay answers when writing in their own discipline and when writing on a more generally relevant topic, and that any such qualitative difference would operate to their advantage in being tested on a SAP rather than a GAP test?

The objection that the specificity in the ELTS does not accurately reflect discipline specificity is investigated indirectly through the qualitative studies in Chapters 5 and 6.

1.5. Expectations ELTS M2 must fulfil

ELTS M2 must fulfil the same expectations as all other writing tests. As a component of a large-scale testing service, the results of which are used to make major decisions affecting the future lives and careers of testees, these expectations operate upon M2 with considerable force. Expectations include the psychometric ones placed upon all language tests, as discussed in Chapter 1, Section 2 and upon all writing tests, as discussed in Chapter 2, but also expectations of construct validity which relate to writing as a construct, as detailed in Chapter 1, Section 1 and to the construct of divisible proficiency on the skills dimension (Chapter 3, Section 1) and the discipline dimension (Chapter 3, Section 2). This latter expectation is additional to those normally demanded of writing tests.

A writing test needs to balance the reality of a test environment with the simulation of a writing purpose and audience other than that of the test. It needs to allow enough space for each writer to show what she or he can do, yet be constrained enough to ensure stable scoring. It needs to not only use valid tasks but also use valid scoring criteria. It needs to be practical for writers, raters, and score consumers, but not at the expense of validity or reliability. The backwash it creates should be

beneficial, in terms of the amount and kinds of writing taught to and valued by testees and potential testees.

In the context of ELTS M2, the writing test, as a direct test of writing performance, needs to mirror (as closely as is possible in the testing situation) the construct of writing as a process of composing, that is, it needs to be psychologically real. As a direct test of writing performance in academic contexts, it needs to mirror as closely as possible the actual writing successful applicants will do on their academic courses, that is, tasks need to be, or at least to appear to writers to be, authentic.

The final rationale for the development of a test of the complexity of ELTS must be that it provides a fairer measure of a testee's language proficiency than the test or tests which it replaces. Thus the rationale for a SAP writing test such as M2 must be that it yields information which corresponds more closely to what the testee can actually do in regard to the writing required in British postgraduate education than a GAP writing test. The implication is that it should provide more information altogether since, if the SAP and GAP writing tests each yield scores in terms of a single number or a general description which cannot be interpreted in terms of the specific writing requirements of the course of tertiary study for which the testee has applied, much of the potential information of a SAP writing test is lost. The information must be at least as reliable and valid as the information yielded by GAP writing tests, and ideally more so.

The study of the various literatures in Chapters 1 to 3 has made it possible to know what the expectations for a SAP writing test should be, and also to know what has been done to date in terms of satisfying those expectations. Thus it is possible to say, on the basis of the preceding chapters, that it is exceedingly difficult for any writing test to satisfy the expectations we must place on any test being used for major placement

decisions, and that it is likely to be considerably more difficult for a SAP writing test to do so.

Carroll (1978) said.

> *... we will need to ... devise workable instruments to measure how far applicants can meet (specific) demands. We must, in doing so, effect a demonstrable improvement on the present system and ensure that the new test itself is capable of continual monitoring and improvement. (p 4)*

The study which follows attempts an assessment of the performance of ELTS M2 in these terms and in terms of the implicit validity claims it makes, but also attempts to go beyond this in considering how any SAP writing test can meet these expectations.

2. DESIGN OF THE STUDY

The study which follows is a validation study in Cronbach's (1971) terms.

> *When validating a decision-making process, the concern is with the question: What is the pay-off when decisions are made in the proposed way, and how does this compare with the pay-off resulting when decisions are made without these data? (p.448)*

Since any SAP writing test will be more difficult to construct and score than a more general writing test because of the increased commitment to design of components and the need for specialist informants, it is essential to establish that there is compensating pay-off in improved information. It is not enough to show that a SAP writing test and a GAP writing test provide equally reliable and valid information. If it could be shown empirically that ELTS M2 provides improved information over a GAP writing test, the objections to M2, which are not empirically based, would be overcome and the development of further SAP writing test instruments would be encouraged. If an improvement in information was not shown, the value of SAP writing tests would be in question.

2.1. Questions and Hypotheses

2.1.1. Main research question

What are the effects on writing scores of overseas non-native postgraduate students at British universities when these testees are asked to write on topics closely related to the content of their own academic discipline (a 'specific academic purpose' - SAP - topic) compared to a topic accessible to all members of the university community (a 'general academic purpose' - GAP - topic)? How can these effects be accounted for?

The hypotheses for the main research question were:

1. There would be no significant differences between the scores assigned to the writing of non-native postgraduate students at British universities when writing on SAP topics and scores assigned to the same students when writing on a GAP topic.

   Differences were investigated through analysis of variance and post hoc Scheffe tests on mean scores, and through score correlations. The significance level was set at .05, using a two-tailed test for correlations, i.e., the probability of a positive result occurring by chance was 5 in 100.

2. Scores assigned to the same subjects for two 'parallel' SAP questions would share more common variance than scores assigned to the same subjects for one SAP question and one GAP question.

   There were two specific aspects to the hypothesis:

a. two SAP questions would exhibit mean scores which would not be significantly different; further, two SAP questions would result in correlations which would enable them to be treated as parallel items, i.e., a correlation of at least .80, 64% of the variance or more in common;

b. mean scores on two SAP questions would be significantly higher than mean scores on a GAP question; further, neither SAP question would share as much variance with the GAP question as they would share with each other: whatever the amount of shared variance for the two SAP questions, the hypothesis would be accepted unless one or both of them shared more variance with the GAP question.

2.1.2. Subsidiary research question

Are scores assigned to answers to essay test questions when scored by the ELTS M2 (first version) procedure adequately reliable for operational and research uses?

The hypotheses for the subsidiary research question were:

1. Single-rater scores resulting from the ELTS M2 (first version) scoring procedure would not be adequately reliable for operational use. The confidence level was set for this study at .80.

2. a) Single-rater scores would not be adequately reliable for research use.

b) By using three raters and combining and averaging scores for each answer, aggregate scores would be obtained which would be adequately reliable for research purposes.

2.2. Subjects

The study began with 126 subjects, of whom 15 were removed from this investigation because of incomplete data. Complete data were available for the remaining 111, who were all postgraduate students. Of these, 103 were Edinburgh University students who matriculated in 1983 or 1984, and 8 were postgraduate students from other universities who attended in-sessional courses at the University of Edinburgh in 1983 and 1984. All the subjects were taking taught Master's degrees except the medical doctors, who were following advanced medical courses at the Postgraduate Board of Medicine at the University of Edinburgh. Table 6.1 shows Modules and overall ELTS scores for the subjects in the sample.

Table 6.1: Breakdown of the sample

Modules and Ns

General Academic	Life Sciences	Medicine	Physical Science	Social Studies	Technology
24	41	11	7	28	0

Overall ELTS Scores

Score	*2*	*2,5*	*3*	*3,5*	*4*	*4,5*	*5*	*5,5*	*6*	*6,5*	*7*	*7,5*
n=	*1*	*1*	*1*	*3*	*4*	*6*	*17*	*28*	*23*	*15*	*7*	*5*

2.3. Measures

Each subject took three writing tests, which are referred to throughout the rest of this chapter, and Chapters 5 and 6, as M2Q1, SAPQ, and GAPQ respectively.

2.3.1. M2Q1

The first measure was the Version 1 Question 1 questions of ELTS M2 (referred to here as M2Q1: GA, LS, etc). In Version 1 the M2Q1: GA question and the M2Q1: SS question are the same, therefore there are five different questions across the six Modules. Each subject wrote on the question intended in the test design to be suitable for her or his specialist discipline/field. Data are only available for five of the six Modules, as there were no subjects with complete data in the Technology Module. The M2Q1: GA and M2Q1: SS questions are treated separately in all the data analyses because (1) we cannot assume that these two groups of students come from the same population simply because they are treated so on this single question (the second question differs on each of these writing tests, and on the second version of the test the two Modules have separate M2Q1s); (2) this quirk of test design provides a fortuitous opportunity to investigate the main research question through another set of data.

The texts and questions are given in Appendix A1. Each of these questions requires the testee to refer to a short (200-100 words, depending on the Module) text which she or he has already read as part of the study skills sub-test of the ELTS, to relate this text to the question, and to bring personal experience and/or opinions into a written response. This kind of response is described as 'divergent' (British Council/University of Cambridge Local Examinations Syndicate memo, 1985). Although many testees will be familiar with the topics, there is no means of ensuring that every testee is familiar with the subject matter of the text and/or the question. The time available for this question is 25 minutes.

2.3.2. SAPQ

The second measure was a 'parallel' question to each M2Q1 question. As we saw in Chapter 2, Section 3.4, there are few guidelines for the design of writing test tasks, and (in Chapter 2, Section 4.5.) fewer for the design of second language writing test tasks. Work conducted by Rose (1980), Johns (1981), Wall (1981), Weir (1983), Horowitz (1986b, and forthcoming) has suggested some parameters of faculty expectations of student responses to tasks in their disciplines, and has examined some features of limited numbers and cross-sections of disciplinary writing test tasks, but we do not as yet have a set of clear models for the design of SAP writing test questions.

Essay test questions from a range of disciplines and subjects at several British universities were collected and reviewed in the search for a model, but these all assumed that every testee answering the question had attended a lecture course and done considerable reading in the specific topic of the question, and had assimilated a depth of content relevant to it. This finding is supported by Wall (op cit) and Weir (op cit). A small study based on a faculty survey by this researcher at the University of Edinburgh, discussed in more detail in Chapter 6, Section 2, found that faculty had few clear criteria for the design of the writing test tasks they set, other than "clarity" or "unambiguousness", and that faculty expectations of responses relate to evidence of content mastery and assimilation for purposes of application in solving new problems.

None of the foregoing is very helpful in the context of task design for ELTS M2, where testees are located all over the world and have followed quite different courses of study up to the point of testing. Questions must be based in what the test constructors can be confident that the writers will all know, which essentially means in an input text. The requirement that testee writers should bring in their personal experience or opinions, used in M2Q1, is questionable in this regard. For this

reason, this requirement was excluded from the design of SAPQ questions; the exclusion also permits a clear distinction between SAPQ and the GAP question on this parameter.

As noted earlier, there is no information available as to the basis for the design of M2Q1 by the ELTS team - not even any statements as to what characteristics are seen as making the M2Q1 questions discipline-specific. Therefore, design of 'parallel' SAP questions had to be carried out without any formal parameters. Although the literature reviewed earlier was of some help, the principal design criteria were; (1) that the 'parallel' questions should be based in the texts in the Source Booklet; (2) that the 'parallel' questions should seem as similar to M2Q1 questions as possible in most ways (excluding the personal experience requirement, as noted above). The second SAP writing test was constructed to consist of six questions, referred to here as SAPQ: GA, LS, etc , based on the same source booklet, each intended to be matched with the M2Q1 question in the same Module (different questions were prepared for SAPQ GA and SAPQ: SS). The same time constraints as for M2Q1 applied. The only Module in which testees would need to bring in background knowledge in order to satisfactorily fulfil the task was Physical Sciences, where the text used as the input provides no basis from which to draw material for a text-based answer to any meaningful writing question. The questions appear in Appendix A2. Each subject wrote on the question intended in the test design to be suitable for her or his specialist discipline/field.

2.3.3. GAPQ

The third measure was a question intended to be discipline-free, that is, a GAP question, referred to throughout the study as GAPQ. A GAP question is perceived as the kind of question which will generate expository discourse without calling on any background knowledge; questions of the GAP type have been described by, for example, Henning (1986), Jones (1986), Bridgeman and Carlson (1983), Fein (1980), Williams (1982) and

Howe (1980), and are commonly used for placement in EAP programmes in British and American universities. The GAP question used here consisted of a short article from the Guardian, and a request for a personal response by each writer. All subjects wrote on the same GAP question. Ten minutes were allowed for reading the text since this had not been seen previously, followed by 25 minutes writing time, i e , the same amount of time as given for ELTS M2 and the SAP questions. The GAP question appears in Appendix A3.

2.4. Procedures

The testees wrote their M2Q1 essays as part of a sitting of the ELTS test as a matriculation requirement of the University of Edinburgh: they were told to spend about 25 minutes on this question, although they could have spent longer by sacrificing part of the time for the second (not investigated) question. Testees wrote their M2Q1 answers immediately after taking M1, during which they had read the input text and answered ocmprehension question on it. The same testees wrote their 'M2 parallel' essays (i.e., SAP: GA, etc) between six and nine weeks later; on this occasion they were allowed 25 minutes, after 5 minutes to refamiliarise themselves with the input text. On the same occasion they wrote the GAP essay, for which they were allowed 35 minutes, including 10 minutes for reading the brief input text.

Clearly, there may have been some development of writing skill in the intervening period, but as Co-ordinator of the in-sessional EAP programme for University of Edinburgh overseas students this researcher worked closely with supervisors to determine the language tuition needs of this group at various points in the university year, in relation to their language use on their courses, and all indications were that most courses require little writing before the Christmas vacation. Criper & Davies (1986) found that the subjects in their test/re-test sample did not increase their scores on M2 over an eight-month period. McKenna, Clark &

Zorn (forthcoming) with first language writers, and Varonis (personal communication) with ESL writers, have found that writers often make little or no progress in the first year of university study.

The data collection exercise resulted in three writing samples from each testee: two on different SAP questions on the same text, and one on a GAP question on a general academic purpose text.

The scoring procedure used for the investigation of these research questions was the first version scoring procedure for question 1 of M2, which was the scoring procedure in operational use during the period of the investigation. In the first version scoring procedure, the rater incorporates into the score on the first question a judgement of the quality of the second question: this stage of the procedure was not included in the study, since only question 1 was under investigation.

Three experienced M2 raters, all also qualified and experienced EFL teachers, each scored each of the three essays. None of the raters was familiar with any of the testees. Identifying information was removed from the subjects' papers, and the order of appearance of the subjects' answers in each of the three sets was randomised, although each rater's set of papers was in the same order as those of the other raters. The scoring procedure as it pertains to M2Q1 is discussed in detail in Chapter 5, and a full copy of the scoring instructions appears in Appendix B1; the essay scale also appears in Appendix B1. The raters received no special training for scoring for this study, in order that the reliability question could be investigated.

3. THE STUDY: SUBSIDIARY RESEARCH QUESTION

It was necessary to complete the investigation of the subsidiary research question prior to the commencement of the main investigation. If scores of criterion reliability for use in the research of the main question could not be obtained, the investigation of that question could not go ahead.

It was shown in Chapter 2 that the question of reliability of essay scores is a complex one. In the context of this study, there were two aspects to the subsidiary research question: are the scores used in the study sufficiently reliable to permit their use for research into the question of the difference, if any, between testee performance on SAP as opposed to GAP essay questions?; and, are scores as used in operational administrations of M2 adequately reliable to justify their use in making decisions about scholarship applicants' likelihood of succeeeding in writing in real academic settings? As stated above, the scoring procedure used for this investigation was the first version scoring procedure (1980 - 1985).

For the investigation of the operational reliability of the first version scoring procedure, it was hypothesised that the scores of single raters would not meet the operational reliability level of .80. The level of .80 reliability for an operational writing test score conforms to the recommendation of Breland & Jones (1982). It is this reliability figure which it is critical for M2 to meet, since the test as used operationally uses only a single rater.

For the investigation of the reliability of the scores used in the investigation of the main research question the hypothesis was that these scores, i.e., the aggregate scores of three experienced raters, would meet the requirement of a reliability level of .80. Because it was predicted that investigation of the data for the main research question would

involve looking at relatively small differences between scores it was essential to achieve this level of score reliability; poor reliability would make it impossible to distinguish the effects of error variance from the effects of true variance. However, reliability levels achieved in this way are appropriate only for research purposes, since the test cannot be scored operationally by three raters. An aggregate score reliability of .80 assumes a single-rater reliability of below .65, a pessimistic assumption for an operational test.

3.1. Data analysis

First, to establish operational reliability, Pearson product-moment correlations were computed for each rater with each other rater on each set of essays. As Diederich (1974) states, this estimate of the amount of agreement between two raters on the same essays is our best measure of the reliability of a single rating. thus these uncorrected correlation coefficients represent a range of single-rater reliabilities such as those to be expected in practice. For each question an average single-rater reliability was calculated by averaging the three obtained single-rater reliabilities. The average single-rater reliability can be viewed as the likely operational reliability level of the test.

Second, the reliability of the scores used for the research of the main research question was established by (1) summing the judgements of the three raters for each answer by each student to obtain aggregate scores, (2) applying the Spearman-Brown correction to estimate the reliabilities of the aggregate scores.

3.2. Results

Table 6.3.1. shows inter-rater reliabilities, i.e., estimated single-rater reliabilities, and average single-rater reliabilities for each question:

Table 4.3.1.: Inter-Rater Correlations

n = 111	M2Q1	SAPQ	GAPQ
Raters 1/2	.686	.655	.602
Raters 1/3	.773	.657	.638
Raters 2/3	.713	.676	.734
All Raters (average single-rater reliabilities)	.724	.663	.658

Inter-rater correlations, i.e., single-rater reliability estimates, ranged from .60 to .77 for different raters on the same question: the average reliability of a single score in this study is .682. This is the probable reliability level for M2 scores on the first question in operational practice.

The estimated aggregate score reliability for each question using the Spearman-Brown correction is shown in Table 4.3.2.:

Table 4.3.2.: Aggregate Score Reliabilities

	M2Q1	SAPQ	GAPQ
Three Rater Reliability (Spearman-Brown)	.887	.855	.852

The average aggregate score reliability, that is, the average reliability of scores to be used in the investigation of the main research question, was .865.

3.3. Discussion

On no occasion does the single-rater reliability reach the criterion level for operational reliability of .80. Thus the first hypothesis, that operational scores do not reach the required reliability level, appears to

be confirmed by these data. An average single-score reliability of .689 indicates that any two raters are only agreeing about 36% of the time on the value of a piece of writing: for a single-rater test such a small amount of agreement is of some concern. The general level of inter-rater reliability compares favourably with, for example, Cast's 49 (1940), Hartog's .51 (1941), Britton et al's .51 (1966) and is close to Wiseman's .70 (1949). Finlayson (1951), reports .73 and Jacobs et al (1981) obtained .74 for single-rater reliability, figures which are only obtained on two of nine occasions here. The score range is generally a little lower than, but does not differ dramatically from, the range of single rater reliabilities reported by Carlson et al (1985). However, the instrument which Carlson et al were validating does not use single-rater scores operationally, but scores from at least two raters, with a third rater when the two scores are discrepant. the scores used operationally consistently reach .80.

Certainly such unreliable scores could not be used in the investigation of the main research question.

It can be seen that inter-rater reliabilities on M2Q1 were consistently higher than on SAPQ, and generally higher than on GAPQ. It is difficult to know whether this is a pattern, or a chance of these data. If there is a pattern, it might be explained by the fact that all the raters had had considerable experience rating M2 papers previously, whereas the SAP and GAP questions were new to them. Anderson (1960) speaks of the practice effect in essay scoring:

> *Somehow or other the standard essay of the marking schedule becomes assimilated and absorbed in the standards customarily and naturally adopted by the marker. (p. 101)*

It might follow from this that raters were less self-consistent in rating SAPQ and GAPQ. Without an intra-rater reliability study it is impossible to take this further. The GAPQ question generated somewhat more erratic

reliabilities than the other two questions: again, there are no data to permit a consideration of why this was.

In regard to the second hypothesis, clearly the single-rater reliability is not at criterion level for use in the investigation of the main research question: therefore hypothesis 2a is confirmed. The Spearman-Brown correction is based on the assumption that reliability is partly a product of test length: thus the reliability of the aggregate score from three raters is greater than the average of their separate reliabilities. Applying the Spearman-Brown correction the reliability levels for three raters ranged from .852 on GAPQ to .887 on M2Q1; the average reliability of the aggregate score for three raters on one essay was .865. Thus hypothesis 2b, that aggregate scores would reach a reliability of .80 or above is confirmed in each case, although somewhat narrowly in the case of GAPQ. The investigation of the main research question would be able to go ahead using these aggregate scores

Investigation of the subsidiary research question had included the calculation of the correlations between all raters on all questions. . If raters are scoring consistently, we should see a pattern of higher correlations for different raters on the same question than for different raters on different questions, and for the same rater on different questions than for different raters on different questions. That is, we should expect raters' judgements of the actual quality of the same essays on the same question to share a good deal of variance; we should expect one rater to exhibit an individual set of responses which is to some extent separate from the actual quality of the essays she is rating; and when judgements being compared have neither the same question nor the same rater in common, but only the same writer, the amount of shared variance should be expected to be quite small. The expectation that there will be higher correlations across raters for the same question than within the same rater for different questions suggests an underlying construct in which: (1) there are some common features of 'writing

proficiency' which skilled raters can recognise in a writing sample; (2) different tasks call on different aspects of the writer's 'writing proficiency'; (3) skilled raters can recognise and agree on the appropriate aspects of a writer's 'writing proficiency' which are called for by a particular question. The expectation that when neither question nor rater are held in common there will still be some common variance across ratings suggests an underlying construct in which writers are viewed as possessing some unique and constant 'writing proficiency', however small a part of the total, which will appear in all their writing regardless of the task and of who rates their performance.

In correlations among all raters on all questions, then, we can consider three kinds of interactions. type 1, correlations for different raters on the same question (classical inter-rater reliability); type 2, correlations for the same rater on different questions, type 3, correlations for different raters on different questions. Our expectation is that correlations would be progressively weaker from type 1 to type 3 interactions. Such a pattern has been found in other studies For example, Vernon & Millican (1954) reported an inter-question reliability of .25 for different markers and .36 for the same marker; Godshalk, Swineford & Coffman (1966) found average inter-question correlations for different readers of .22 to .30, and for the same reader of .36 to .41. Carlson et al (1985) report only inter-question correlations between topic types, using the holistic score, i.e. the score of two or three raters combined. Their correlations were between .66 and .73 regardless of topic type but they do not report single-rater correlations across questions. Pollitt (personal communication) found that in a study of five different writing tasks by the same writers, uncorrected inter-question correlations were around .3 for the same marker. Pollitt did not investigate correlations across markers.

In Table 4.3.3., type 1 correlations are shown in **bold face**, type 2 correlations in *italic*, and type 3 correlations are in normal face:

Table 4.3.3.: Rater/Question Correlation Matrix

		Rater 1			Rater 2			Rater 3		
		M2Q1	SAPQ	GAPQ	M2Q1	SAPQ	GAPQ	M2Q1	SAPQ	GAPQ
Rater 1	M2Q1	1.00								
	SAPQ	*.756*	1.00							
	GAPQ	*.499*	*.711*	1.00						
Rater 2	M2Q1	.686	.612	.433	1.00					
	SAPQ	.312	.655	.406	*.331*	1.00				
	GAPQ	.487	.687	.602	*.505*	*.404*	1.00			
Rater 3	M2Q1	.773	.612	.486	.713	.185	.107	1.00		
	SAPQ	.312	.657	.416	.284	.676	.404	*.255*	1.00	
	GAPQ	.488	.280	.638	.498	.219	.734	*.474*	*.411*	1.00

Table 4.3.3. shows us, as discussed above, that inter-rater reliability, i.e., type 1 correlations, for these data varied between .602 and .773. that is, raters were agreeing on about 36% to 60% of the variance in the same essays. Table 4.3.3. shows that type 2 correlations range between .255 (rater 3, SAPQ/M2Q1) and .756 (rater 1, SAPQ/M2Q1), with an average of .483. Type 3 correlations range between .107 (rater 2 GAPQ/rater 3 SAPQ) and .687 (rater 2 GAPQ/rater 1 SAPQ), with an average of .409. We see, therefore, the same pattern of support for the underlying constructs detailed in the preceding paragraph as have been found in other studies.

While in general we find support in Table 4.3.3. for the constructs discussed above, we also find that not all raters were similarly affected by the rating of different questions. Rater 1 rated the two SAP questions very similarly; his ratings of SAPQ and GAPQ also correlate highly (.711). Rater 3, in contrast, does not rate any of the three sets of essays very similarly, and the SAPQ/M2Q1 correlation is very low at .255 (only 6% of the variance). Rater 2 performs more like rater 3 than rater 1. When we look at the ratings both across raters and across

questions, we see that SAPQ/M2Q1 and SAPQ/GAPQ resulted in a wide range of correlations across raters, whereas M2Q1/GAPQ resulted in a narrow range between .433 and .498. This data configuration suggests that GAPQ and M2Q1 may be somewhat weakly but consistently related. The correlations between SAP questions for different raters seem not to exhibit an easily interpretable pattern.

3.4. Implications

The range of correlations displayed in Table 6.3.4 , representing as it does the responses of several raters to the same and different tasks, can be expected not to differ too much from the range of responses of raters of ELTS M2, in its original version scoring procedure, in practice. Indeed, it may represent the upper limit of reliability of operational M2Q1 scores, since the raters used in this study were all highly qualified EFL teachers, possessing at least a Master's degree in ELT or Applied Linguistics, and all had had considerable teaching experience both in Britain and outside it. They were chosen as being the best raters, because of the extreme importance for the investigation of the main research question that scores be as reliable as possible. With raters rating in isolation from each other, coming from widely different backgrounds and working in very different contexts, holding fewer qualifications, having fewer years teaching experience and less experience of different language levels, cultures and discourse communities, we can expect inter rater reliabilities much lower than those that were found in this study - probably of the order of those found by Diederich et al (1961) in rather similar circumstances (described in Chapter 2, Section 3), where the median inter-reader reliability was .31.

The circumstances in which M2 is administered place severe constraints on the test operationally. Some raters have very little practice, rating only one or two papers each month. Many raters also deal with a very limited sample, seeing only essays written by writers from one

country/language background for several years on end. Even with careful scoring procedures, it is very difficult for such raters to stay on-scale. Many different raters rate the same question; and the same rater rates (in 1986) twelve different M2Q1 questions. Yet scores are expected to be equivalent regardless of who did the rating and what question the testee was assigned, and these scores are treated and reported as if they are indeed equivalent.

The lack of precise guidance; the imprecision of the criteria and performance descriptions; awareness of the tremendous constraints imposed by the operational context had already caused concern among British Council English Language Officers and other ELT professionals involved with ELTS. A good deal of anecdotal evidence had already accumulated when this study was carried out. For this researcher, practical experience with the first version scoring procedure had confirmed that it put great demands on the rater's professional knowledge and experience, demands which it is difficult to be certain are met in operational contexts.

While the aggregate scores had been shown to be adequately reliable, it was clearly impractical to propose such a system of multiple-marking for the M2 context. The single-rater system would have to continue, at least into the foreseeable future. The study of the subsidiary research question had highlighted an existing awareness of the urgent need for a new scoring procedure and a method of rater training which would improve the reliability of scores of single raters.

4. MAIN RESEARCH QUESTION

The main research question asked what would be the effects on writing scores of non-native postgraduate students at British universities when they are asked to write on topics closely related to the content of their own academic discipline (a 'specific academic purpose' - SAP - topic) compared to a topic accessible to all members of the university community (a 'general acdemic purpose' - GAP - topic). The intention for the study of this question was that attempts would be made to suggest explanations for any observed effects.

There were two hypotheses for the main research question. Although the claim for ESP testing, as was seen in Chapter 5, is that students will be advantaged by being tested through material relevant to their specialist knowledge, Weir (1983), in the only in-depth study of the impact of ESP testing to date, concluded that:

> *We were unable to produce any conclusive evidence that students were disadvantaged by taking tests in which they had to deal with texts other than those from their own subject area. The case for a variety of ESP tests therefore remains unproven. (p. 550)*

In the survey of the literature no empirical studies directly relevant to the main research question had been found. It was therefore felt that there was insufficient previous research evidence to permit a prediction of whether testees would exhibit differential performance on a SAP question or a GAP question, and therefore a hypothesis of no difference was selected. The first hypothesis, that there would be no significant differences between testees' scores on SAP test questions and GAP test questions, would be accepted if both interactions, M2Q1/GAPQ and SAPQ/GAPQ, demonstrated a relationship between scores significant at $p \leq .05$ on a two-tailed test, i.e., probability less than 5 in 100 cases that the pattern is due to chance, and if no significant differences were found among mean scores.

The hypothesis of no difference would be considered to be neither confirmed nor rejected if only one correlation, M2Q1/GAPQ or SAPQ/GAPQ showed the hypothesised pattern of no difference. It would also be considered to be neither confirmed nor rejected if there was not a clear pattern in one direction across Modules: that is, if one or two Modules showed significant differences between SAP questions and the GAP question but others did not, or if all Modules showed significant differences but not in the same direction, the hypothesis could not be either rejected or confirmed.

The second hypothesis arises from the discussion of task variables in Chapter 2, Section 3 4. It was seen there that in discussions of score validity in previous studies it has sometimes been claimed that the topic a testee is assigned makes no or insignificant difference (e.g., Carlson et al, 1985). The fact that each of the ELTS Modules exists (in 1986) in two versions, and that the writing tasks differ from version to version, is based on the assumption that two writing tasks on the same reading material function as parallel forms of the test, i.e., that testees are neither advantaged nor disadvantaged by the random assignment of one form or the other. A similar assumption that all writing test questions are equivalent, without the need for pre-testing to measure equivalence, seems to underly a number of studies, which sometimes do not report whether in their data collection all testees wrote on the same topic on all occasions (e.g., Fein, 1980; Kroll, 1982, Robinson, 1985).

However, there is growing concern at the present time that the impact of the actual topic on essay tests is greater than has been acknowledged, and that the reasons for this are little understood. Jacobs et al (1981) found that score reliability (reliability of scores of the same students on more than one topic) was lower than rater reliability, and that "student performance does indeed vary from topic to topic" (p. 73); White (1985) reports data from the California State University English Equivalency Examination to show that two topics requiring different types

of writing (expressive/explanatory) when written by the same testees produced very low correlations: as years passed and the types of writing required by the two topics became more similar the correlations became higher. In this second phase of the study, it was predicted that scores on M2Q1 and SAPQ would be more closely related to each other than the score on either of them would be to GAPQ. Specifically, it was predicted (hypothesis 2a) that the two SAP questions, M2Q1 and SAPQ, would result in correlations which would enable them to be treated as 'parallel' items. The criterion for this was set at $r \leq .80$, i.e., a shared variance of 64% or more, below which hypothesis 2a would be rejected. To be truly parallel, the two SAP questions should also generate mean scores which are not significantly different. Further, it was predicted (hypothesis 2b) that neither M2Q1 or SAPQ would show an equally strong relationship with GAPQ. Hypothesis 2b would be rejected if the amount of variance shared by GAPQ and either M2Q1 or SAPQ exceeded the amount shared by M2Q1 and SAPQ.

These hypotheses would be considered to be neither confirmed or rejected unless there was a pattern of the predicted relationship from Module to Module; they could not be rejected if only isolated or non-systematic instances were found.

4.1. Data Analysis

As discussed in Section 2.4 of this Chapter, the scores for each subject on each question by all three raters were combined and averaged to give an aggregate score. The investigation of the subsidiary research question in the previous Section had ascertained that these aggregate scores met the criterion level of reliability for use in the main research study. The reliability levels for the scores used here are:

M2Q1: .887 SAPQ: .855 GAPQ: .852 (n = 111)

Following the procedure used by the Edinburgh ELTS Validation Study (Criper and Davies, 1986), the data were analysed for the whole group and for each Module separately. Means and standard deviations for each question were obtained for the subjects considered as one group Means and standard deviations were also obtained for the whole group for ELTS overall score ('ELTSOA') and for ELTS score excluding M2Q1 ('ELTS-M2'), to provide comparative measures for the group. Analyses of variance were carried out, with post-hoc Scheffè comparisons of any significant results. Pearson product-moment correlations and significance levels were obtained for each pair of questions for the whole group, and for each writing test with the two comparative measures.

The subjects were then sorted into sub-groups by the Module into which they fell. The n in the Physical Science sample (n=7) is very small and therefore the figures may be suspect: however, they are included for any additional light they may shed on the SAP/GAP issue. The same procedures were carried out for each Modular group as for the whole group, as discussed in the preceding paragraph. Patterns of relationships among the Modular sub-groups were then examined.

There was no *de facto* ground for analysing these data either as a single group or by Modular divisions: that is, there was no pre-existing evidence to say that if they were not analysed separately important behaviours would be masked. Indeed, this is the question which the study was designed to answer. Therefore the data were analysed both as a single set and by Modular divisions for each aspect of the main research study. In this way it was hoped that it would become clear how much, if any, additional information was provided by the more finely-grained analyses. It was also hoped that these finely-grained analyses would reveal areas of the data where further study was called for. In the next two sub-sections the data are analysed, first for the group as a whole, then for the group broken down by Module. Each aspect of the data is

discussed as it is presented, because of the close inferential links between different stages of the data analysis.

4.2. Results and discussion: whole group

The means and standard deviations for the three questions when considered for the whole group are shown in Table 4.4.1.

Table 4.4.1.: Means and Standard Deviations

	$\bar{x}$	sd
M2Q1	5.477	2.106
SAPQ	5.685	1.688
GAPQ	4.919	2.202

The mean scores for the two SAP questions, M2Q1 and SAPQ, are quite similar, while the mean score for GAPQ is considerably lower. For both SAP questions the mean score is around band 5.5, which on the ELTS M2 rating scale (Appendix B1) falls between 'Modest Writer' (band 5) and 'Competent Writer' (band 6), while for GAPQ the mean score is slightly above band 4.5, which falls between 'Marginal Writer' (band 4) and 'Modest Writer' (band 5). In terms of the description attached to the performance this is quite an important difference; it is also important in terms of the likelihood of an 'average' testee, i.e., an individual testee who scores at or around the mean, being found acceptable for a postgraduate course of study in Britain.

Analysis of variance indicated significant difference ($p < .001$) among these means (Table 4.4.2.):

Table 4.4.2.: ANOVA for Whole Group Means

Source of variance	SS	d.f.	MS	F	p
Between groups	34.829	2	16.915	8.458	< .001
Within groups	659.928	330	1.999		

Post hoc Scheffè (α = .05) indicated the following significant differences between means: (Table 4.4.3.):

Table 4.4.3.: Post-hoc Scheffè Comparisons of Whole Group Means

t_{crit}	t_{obs}		p
2.461	M2Q1/GAPQ	2.937	.05
	SAPQ/GAPQ	4.032	.01

One way analysis of variance confirmed the impression that there were significant differences among the means, and the Scheffè post-hoc comparisons located these between the means of both SAP questions and the GAP question, although the difference was greater for SAPQ/GAPQ than for M2Q1/GAPQ.

Table 4.4.4. shows means and standard deviations for the whole group on ELTSOA (overall ELTS score, which includes a single-rater M2 score) and ELTS-M2 (ELTS overall score excluding the M2 score):

Table 4.4.4.: Comparative ELTS Means

	$\bar{x}$	sd
ELTSOA	5.696	.871
ELTS-M2	5.696	.910
M2Q1	5.477	2.106
SAPQ	5.685	1.688
GAPQ	4.919	2.202

It can be seen that these mean scores are very close to the score for SAPQ and quite close to the score for M2Q1, but considerably higher than the GAPQ score. Whatever ELTS as a whole is measuring, it would appear that the SAP questions have more in common with it than the GAP question does. Further, for these subjects it would appear that the inclusion of the writing test in the ELTS makes no difference to their overall scores. It must be remembered that the M2 scores which are included in ELTSOA

and excluded from ELTS-M2, i.e., which form the basis of the relationship between the two means in Table 4.4.4., are not the scores used to generate the aggregate scores for this study. Nevertheless, if the argument for the inclusion of a direct writing sample in ELTS is one of divisible proficiency, there is no evidence in Table 4 4.4. to support such an argument. One way analysis of variance was carried out to determine whether there was significant difference among the five means (Table 4.4.5.):

Table 4.4.5.: ANOVA for Whole Group Means with Comparative ELTS Means

Source of variance	SS	d.f.	MS	F	p
Between groups	48.857	4	12.214	4.408	< 01
Within groups	1523.838	550	2.771		

Post hoc Scheffè (α = .05) indicated the following significant differences between means (Table 4.4.6):

Table 4.4.6.: Post-hoc Scheffè Comparisons for Whole Group Means with Comparative ELTS Means

t_{crit}	t_{obs}		p
3.09	ELTSOA/M2Q1	4.38	< .01
	ELTSOA/GAPQ	15.54	< .001
	ELTS-M2/M2Q1	4.38	< .01
	ELTS-M2/GAPQ	15.54	< .001

It can be seen in Table 4.4.6 that SAPQ is the only one of the three writing tests which does not differ significantly from the comparative ELTS means. On the basis of this more detailed analysis of the data, it would appear that there would be no justification for the inclusion of SAPQ within the ELTS battery, since it is contributing no additional information. However, M2Q1 does make a significant contribution, and the contribution is even greater for GAPQ (note, however, that the operational M2 does not make a significant contribution). If the inclusion of a writing test is to be justified from the basis of a divisible skills

argument, the GAP test would appear to be the most useful inclusion, since it differs markedly from ELTS-M2. On the basis of this argument the fact that GAPQ is a 'general academic' rather than a 'specific academic' purpose writing test is not important. On the other hand, if the inclusion of a writing test is to be justified on the basis of divisible specialist proficiencies (the SAP/GAP question which is at the heart of this study), these data are somewhat equivocal, since the two supposedly 'parallel' SAP questions are performing quite differently in relation to the comparative ELTS means. M2Q1 appears to be more like GAPQ than SAPQ. Means, however, obscure the performance of individuals, and we must look at the correlational data to understand how subjects are performing.

The uncorrected Pearson product-moment correlations between the aggregate scores for the three questions are shown below the diagonal in Table 4.4.7. Correction for attenuation (Nunnally, 1978) was applied to both tests in each interaction, using the Spearman-Brown corrected reliabilites for aggregate scores, to estimate how much greater the correlations between the tests would be if each was perfectly reliable: the corrected correlations are given above the diagonal in Table 4.4.7.

Table 4.4.7.: Uncorrected and Corrected Correlations of Aggregate Scores [1]

	M2Q1	SAPQ	GAPQ
M2Q1	1.00	.584	.616
SAPQ	.512	1.00	.582
GAPQ	.539	.497	1.00

[1] All correlations are significant at p< .001.

All the correlations in Table 4.4.7 are significant at p< .001 and thus for the data considered for the whole group the first hypothesis, that there would be no significant differences between M2Q1/SAPQ scores and

the GAPQ score, must be accepted. However, even when corrected for attenuation the actual correlations are not very large and are only accounting for 33% to 38% of the score variance. The two SAP questions do not correlate at .80 or greater, and therefore hypothesis 2a must be rejected: for the group considered as a whole, M2Q1 and SAPQ apparently do not function as parallel forms. Further, the correlations for each of them with GAPQ are not very different than their correlations with each other, and in fact the correlation between M2Q1 and GAPQ is greater than that between M2Q1 and SAPQ: therefore hypothesis 2b must be rejected.

Table 4.4.8. shows the correlations of M2Q1, SAPQ and GAPQ with ELTS overall score (ELTSOA) and ELTS without M2 (ELTS-M2). To arrive at these correlations the reliabilities of the three writing tests have been corrected using the formula for correction for attenuation for one of a pair of tests (Nunnally, 1978) and the reliability levels for the aggregate scores: ELTS scores could not be corrected because no basis was available to determine their attained reliability. To avoid confusion, Table 4.4.8 shows correlations based on partially-corrected reliabilites above the diagonal and the ELTSOA/ELTS-M2 correlation based on uncorrected scores below the diagonal: partially-corrected correlations for the three writing tests with each other are also included in Table 4.4.8.

Table 4.4.8. shows that, while the correlation between ELTS and ELTS-M2 is very high, it is not at or close to 1.00, which might have been expected given the exact coincidence of the mean scores. A small but potentially interesting proportion of the variance remains unaccounted for. This may, of course, be simply error variance, since we have no exact idea of the reliability attributable to either score. It can further be seen in Table 4.4.8. that ELTSOA shares more variance with M2Q1 than with either SAPQ or GAPQ, although the relationship is stronger for GAPQ than for SAPQ. The pattern is slightly different, and less strong, for ELTS-M2, where ELTS-M2 and GAPQ share most common variance: the

difference for SAPQ is quite marked in this case, although again the correlations are all significant at p< .001.

Table 4.4.8.: Whole Group Correlations with Comparative Scores [1]

	ELTSOA	ELTS-M2	M2Q1	SAPQ	GAPQ
ELTSOA	1.00	*	.616	.498	.526
ELTS-M2	.870	1.00	.483	.412	.499
M2Q1	*	*	1.00	.541	.569
SAPQ	*	*	*	1.00	.539
GAPQ	*	*	*	*	1.00

[1] All correlations are significant at p< .001.

It should be noted that corrected correlations indicate the correlations between hypothetical (and unobtainable) 'true' scores, and are valid only in theoretical research: in reporting correlations for operational tests correction for attentuation is invalid and "can lead to dangerously misleading results" (Pilliner, personal communication). It can be seen in Table 4.4.7. that correction for attentuation boosts correlations across the board, but does not change the direction or relative strength of the correlations. Therefore in the tables which follow correction for attentuation is not used. This should be borne in mind when making comparisons between the findings of this study and, for example, the findings of Carlson et al (1985), in which correction for attentuation was employed.

The information from the mean scores appears to be misleading, since it shows most difference between the GAPQ score and the SAP scores, implying a strong relationship between the two SAP scores. Such a relationship is not found in the pattern of correlations, which suggest that individuals are placed differently on the two SAP tests, even though

the means are quite close. There are more similarities between the rank orders of individuals between GAPQ and M2Q1 than between their rank order on M2Q1 and SAPQ: a testee's performance on M2Q1 seems to be better predicted by the GAPQ score than by the SAPQ score.

When analysed for the whole group, these data seem to present some conflicting and confusing patterns of relationships: while the SAP means are significantly different from the GAP mean, all three tests are significantly correlated, yet share less than 40% of the variance. It would seem that the closely related SAP means do not imply that the same subjects are performing similarly on the three tests. Further, although the three tests are highly correlated, the amount of shared variance is not high enough to permit a claim that they are all testing 'the same thing'; nor is there a pattern of differences among amounts of shared variance which would permit an interpretation that one test is testing 'something different' from the others.

It must be remembered that the two SAP means are not mean scores on single questions, but that each was arrived at by combining scores on four or five different questions each representing its question 'type' for one Module (recall that, although there are six Modules in the ELTS, there were no Technology students in this sample, and the M2Q1 question for Social Studies and General Academic was the same.) The justification for treating the scores on these four/five different questions in the SAP tests, which were written by different samples (drawn, by definition, from different populations), as if they are scores on one question is twofold. First, as was stated above, there is as yet no evidence to indicate why this should not be done. Second, in operational terms they are treated as though they are the same. That is, when the English Language Testing Service reports scores on M2, it does not indicate that there might be any difference in score pattern from Module to Module, nor do the recommendations of lengths of tuition make any allowance for such differences (ELTS: Administrators' Manual, no date: received Autumn 1983).

It may be that working with whole group data is obscuring some interesting characteristics of the subjects' peformances. This is a question which the analyses and discussion of results by Module will address.

4.3. Results and discussion: group by Module

In the section which follows the results for each Module are reported and discussed separately; they are then discussed altogether, and some implications considered, in Section 5.

4.3.1. Life Sciences

Means and standard deviations for the Life Sciences sub-group are shown in Table 4.4.9 (LS):

Table 4.4.9 (LS): Means and Standard Deviations

	$\bar{x}$	sd
M2Q1	5.585	1.045
SAPQ	6.293	.698
GAPQ	4.927	1.523

One-way analysis of variance showed that there was significant difference among these means (Table 4.4.10 [LS]):

Table 4.4.10 (LS): ANOVA

Source of variation	SS	d.f.	MS	F	p
Between groups	38.260	38	19.13	5.714	< .01
Within groups	127.22	2	3.348		

Post-hoc Scheffè comparisons (α = .05) showed that the difference occurred in each interaction (Table 4.4.11 [LS]):

Table 4.4.11 (LS): Post-hoc Scheffè Comparisons

(over)

t_{crit}	t_{obs}		p
3.228	M2Q1/SAPQ	4.307	$< .01$
	M2Q1/GAPQ	4.037	$< .01$
	GAPQ/SAPQ	8.380	$< .001$

The interaction is particularly marked in the case of GAPQ/SAPQ. For this sub-group, the SAPQ question yields the highest scores and GAPQ the lowest: M2Q1 appears to mediate between them. GAPQ yields scores which are on average below acceptability level for tertiary education in Britain, while SAPQ yields scores which are on average safely above the acceptability threshold for tertiary education in Britain.

Correlations between the three questions, and significance levels, for the Life Sciences sub-group are shown in Table 4.4.12. (LS):

Table 4.4.12 (LS): Correlations and Significance Levels [1]

(n=41)

	M2Q1	SAPQ	GAPQ
M2Q1	1.00		
SAPQ	.467 p<.01	1.00	
GAPQ	.611 p<.001	.562 p<.001	1.00 .

[1] The correlations for all the Modules are based on the aggregate score, i.e., on scores reliable at above 80, but correction for attenuation has not been applied to Modular correlations.

Examining these data in the light of the first hypothesis, that scores on SAP questions will not be significantly different from scores on a GAP question, we see that this is the case. Each SAP question score is

highly correlated with the GAP score: for the Life Sciences Module, we must accept the first hypothesis.

However, when we turn to the second hypothesis, that SAP question scores would be more highly correlated with each other than with the GAP question score, we see first that the correlation between the two SAP questions does not meet the criterion of r‹ .80: M2Q1 and SAPQ are only sharing about 21% common variance, and we must reject hypothesis 2a. Further, each of the SAP tests shows a higher correlation with GAPQ than they do with each other: we must therefore reject hypothesis 2b.

For the Life Sciences Module we see a pattern similar to that found for the whole group: although the mean score for GAPQ is significantly different from the other two scores, all the tests are significantly correlated. While none of the interactions meet the criterion for parallel forms none of them is different enough from the others to permit a claim that it is testing 'something different'.

It will be useful to know the relationship between the scores on each of the three tests and the overall ELTS score, and the ELTS score without M2: Table 4.4.13 (LS) displays the five means together:

Table 4.4.13 (LS): Writing test means compared with ELTSOA and ELTS-M2

	$\bar{x}$	sd
ELTSOA	5.500	.628
ELTS-M2	5.436	.680
M2Q1	5.585	.611
SAPQ	6.293	.698
GAPQ	4.927	1.523

We see in Table 4.4 13 (LS) that for the Life Sciences sub-group M2Q1 yields scores which are very close to the overall ELTS score for the sub-group. For this sub-group, M2Q1 is contributing very little information to the overall score. As we would expect, this can be seen in the ELTS-

M2 score, which is scarcely changed by the subtraction of M2. On the M2 band descriptors (Appendix B1) a score of 5.5 would be associated with the level 5 band descriptor, 'Modest Writer', as would the GAPQ score which rounds to band 5.0. Only SAPQ yields a score for this group which on average lifts performance into a higher descriptive category, band 6, 'Competent Writer'. On these data, it would appear that the Life Science sub-group is marginally advantaged by SAP questions, with this advantage more marked for SAPQ than for M2Q1. It would also appear that the inclusion of SAPQ rather than either of the other two questions would result in greatest information gain, since SAPQ is least like the overall score.

4.3.2. Medicine

Means and standard deviations for the Medicine sub-group are shown in Table 4.4.9 (ME):

Table 4.4.9 (ME): Means and Standard Deviations

	$\bar{x}$	sd
M2Q1	5.909	1.045
SAPQ	6.000	1.483
GAPQ	5.818	1.401

One-way analysis of variance showed the differences among these means were not significant.

Correlations between the three questions with significance levels for the Medicine sub-group are shown in Table 4.4.12 (ME):

Table 4.4.12 (ME). Correlations and Significance Levels

(n=11)

	M2Q1	SAPQ	GAPQ
M2Q1	1.00		
SAPQ	-.328 NS	1.00	
GAPQ	-.352 NS	.548 NS	1.00

Examining these data in the light of the first hypothesis, we find that there are significant differences between scores on SAP questions and the score on the GAP question. Each of the SAP scores fails to show a significant correlation with the GAP score, although the SAPQ/GAPQ correlation narrowly misses significance at p< .05. M2Q1 and GAPQ show an inverse correlation: that is, those who scored high on M2Q1 scored low on GAPQ and vice versa. The inverse relationship is not significant, however. For the Medicine Module, we must reject the first hypothesis, despite the similarity of the means.

For the second hypothesis, that SAP question scores would be more highly correlated with each other than with the GAP question score, we see first that M2Q1 and SAPQ fail to achieve a correlation of .80 or above. Indeed, the negative correlation between any two writing tests, never mind two questions intended to be parallel is very surprising and must be investigated more closely in later chapters. For the moment we shall note that hypothesis 2a must be rejected for the Medicine sub-group. Hypothesis 2b, that neither SAP question would share more variance with GAPQ than they do with each other, must also be rejected, since M2Q1 and GAPQ show a positive correlation.

Turning to a comparison of the means on the three questions with the means for ELTS OA and ELTS-M2 for the Medicine sub-group, we have the results shown in Table 4.4.13 (ME):

Table 4.4.13 (ME): Writing test means compared with ELTSOA and ELTS-M2

	$\bar{x}$	sd
ELTSOA	6.000	1.294
ELTS-M2	6.045	.723
M2Q1	5.909	1.045
SAPQ	6.000	1.483
GAPQ	5.818	1.401

We see that, for the Medicine sub-group as a whole, the addition of any of the writing tests would make little difference to the overall score. In terms of information gain there is no argument to be made here. For any individual, however, the selection of any one of the three tests rather than the others could have a marked effect: of the eleven subjects in the sub-group there was only one for whom selection of one rather than another of the writing tests did not result in placement into a different score category.

4.3.3. Physical Science

Means and standard deviations for the Physical Science sub-group are shown in Table 4.4.9 (PS): it must be pointed out at this point that there were only seven subjects in the Physical Science sub-group and results must be viewed in the light of this limitation.

Table 4.4.9 (PS): Means and Standard Deviations

	$\bar{x}$	sd
M2Q1	5.428	2.225
SAPQ	5.143	1.676
GAPQ	4.857	1.464

One way analysis of variance showed the differences among these means were not significant.

Correlations between the three questions with significance levels for the three sub-groups are shown in Table 4.4.12 (PS):

Table 4.4.12 (PS): Correlations and Significance Levels

(n-7)

	M2Q1	SAPQ	GAPQ
M2Q1	1.00		
SAPQ	.648 NS	1.00	
GAPQ	.789 p<.05	.893 p<.005	1.00

Examining these data in the light of the first hypothesis, that scores on SAP questions will not be significantly different from scores on a GAP question, we see that this is the case. Each SAP question score is highly correlated with the GAP score: for the Physical Science Module, we must accept the first hypothesis.

For the second hypothesis, that SAP question scores would be more highly correlated with each other than with the GAP question score, we see first that the correlation between the two SAP questions fails to meet the criterion of r< .80: we cannot consider M2Q1 (PS) and SAPQ (PS) as parallel forms and must reject hypothesis 2a for the Physical Science sub-group. As regards hypothesis 2b, we see that the correlation between the two SAP questions is lower than the correlation between either SAP question and the GAP question; in fact the correlation between SAPQ and GAPQ is extraordinarily high and meets the criterion for parallel forms,

although there is no intention nor claim that SAPQ and GAPQ are parallel. We must therefore reject hypothesis 2b.

Table 4.4.13 (PS) shows the comparison between the three writing tests and thescores of this sub-group on ELTSOA and ELTS-M2.

Table 4.4.13 (PS): Writing test means compared with ELTSOA and ELTS-M2

	$\bar{x}$	sd
ELTSOA	5.714	2.903
ELTS-M2	5.886	3.306
M2Q1	5.428	2.225
SAPQ	5.143	1.676
GAPQ	4.857	1.464

We see that without M2 the overall score is slightly higher and the variance is slightly larger: this small sub-group contains a wide range of apparent levels, which have a marked effect because the group is so small. Each of the writing test means would result in a band descriptor at level 5, 'Modest Writer': both ELTSOA and ELTS-M2 would result in an overall score which would be rounded to band 6, 'Competent User'. The inclusion of M2 lowers the mean for the sub-group, which is in accordance with the other results: if it can be generalised that the 'average' Physical Science subject writes slightly less well than she performs other language skills, then the most useful inclusion in the test battery would be GAPQ, which yields the lowest writing score and thus provides the greatest information gain.

4.3.4. Social Studies

Means and standard deviations for the Social Studies sub-group are shown in Table 4.4.9 (SS):

Table 4.4.9 (SS): Means and Standard Deviations

	$\bar{x}$	sd
M2Q1	5.555	.725
SAPQ	4.926	2.618
GAPQ	4.555	1.155

One-way analysis of variance showed that there were significant differences among these means (Table 4.4.10 [SS]):

Table 4.4.10 (SS): ANOVA

Source of variation	SS	d.f.	MS	F	p
Between groups	13 802	2	6 901	9.887	< .001
Within groups	55 186	78	.698		

Post-hoc Scheffè comparisons (α = .05) showed that the differences occurred in all interactions (Table 4.4.11 [SS].

Table 4.4.11 (SS) Post-hoc Scheffè Comparisons

t_{crit}	t_{obs}		p
2.498	M2Q1/SAPQ	24.192	< .001
	M2Q1/GAPQ	38.461	< .001
	GAPQ/SAPQ	14.269	< .001

Correlations and significance levels for the three questions with significance levels for the Social Studies sub-group are shown in Table 4.4.12 (SS):

Table 4.4.12 (SS): Correlations and Significance Levels

(n=27)

	M2Q1	SAPQ	GAPQ
M2Q1	1.00		
SAPQ	.133 NS	1.00	
GAPQ	.444 p<.05	.487 p< 01	1.00

Examining these data in the light of the first hypothesis, that scores on SAP questions will not be significantly different from scores on a GAP question, we see that this is the case. Each SAP question score is highly correlated with the GAP score even though they are not highly correlated with each other: for the Social Studies Module, we must accept the first hypothesis.

For the second hypothesis, that SAP question scores would be more highly correlated with each other than with the GAP question score, we see first that the two SAP questions do not correlate at the criterion level of .80 or above: we cannot consider M2Q1 (SS) and SAPQ (SS) to be parallel forms, and hypothesis 2a must be rejected. Further, the correlation of each of the SAP questions with GAPQ is greater than their correlation with each other, and therefore, for the Social Studies Module we must reject hypothesis 2b.

If we examine Table 4.4.13 (SS), we can see the relationship between the three writing tests and scores on ELTS as a whole. The overall ELTS score for the Social Studies sub-group is band 6, 'Competent User'; this is lowered slightly by the exclusion of M2, which suggests that the operational M2 scores were somewhat higher than the M2Q1 scores used here. The two SAP questions yield mean scores which would be reported

Table 4.4.13 (SS): Writing test means compared with ELTSOA and ELTS-M2

	$\bar{x}$	sd
ELTSOA	5.916	1.005
ELTS-M2	5.854	.886
M2Q1	5.555	.725
SAPQ	4.926	2.618
GAPQ	4.555	1.555

in association with the level 4 band descriptor, 'Marginal Writer'. From this point of view it would be to the average testee's advantage to test on M2Q1 rather than SAPQ, and GAPQ would be least advantageous.

4.3.5. General Academic

Means and standard deviations for the General Academic sub-group are shown in Table 4.4.9 (GA):

Table 4.4.9 (GA): Means and Standard Deviations

	$\bar{x}$	sd
M2Q1	5.261	2.005
SAPQ	5.826	2.167
GAPQ	4.782	2.372

One way analysis of variance showed no significant differences among these means.

Correlations and significance levels for the General Academic sub-group on the three questions are shown in Table 4.4.12 (GA):

Table 4.4.12 (GA): Correlations and Significance Levels

(n=23)

	M2Q1	SAPQ	GAPQ
M2Q1	1.00		
SAPQ	.827 p<.001	1.00	
GAPQ	.739 p<.001	.646 p<.001	1.00

Examining these data in the light of the first hypothesis, that scores on SAP questions will not be significantly different from scores on a GAP question, we see that this is the case. Each SAP question score is highly correlated with the GAP score. For the General Academic Module, we must accept the first hypothesis.

Turning to the second hypothesis, that SAP question scores would be more highly correlated with each other than with the GAP question score, we see that this is the case. The correlation between M2Q1 and SAPQ is not only highly significant, it meets the criterion of r< .80 for parallel forms. On this basis, we must accept hypothesis 2a for the General Academic Module. Further, neither of the SAP questions is correlated more highly with GAPQ than they are with each other, and therefore hypothesis 2b must also be accepted for this Module.

Table 4.4.13 (GA) shows the means for ELTSOA and ELTS-M2 compared with the means of the three writing tests. M2Q1 corresponds most closely with ELTSOA and ELTS-M2; clearly the inclusion of M2 in the overall score lowers this marginally, and this is reflected in the M2Q1 mean score for the sub-group. The average M2Q1 and GAPQ score would be reported as band 5, 'Modest Writer', whereas the average SAPQ score would be reported

as band 6, 'Competent Writer': on this criterion subjects are advantaged by the SAPQ question rather than M2Q1 or GAPQ.

Table 4.4.13 (GA): Writing test means compared with ELTSOA and ELTS-M2

	$\bar{x}$	sd
ELTSOA	5.250	1.141
ELTS-M2	5.350	1 182
M2Q1	5.261	2.005
SAPQ	5.826	2.167
GAPQ	4.782	2.372

4 4. Results and discussion: overall

It is now possible to consider the pattern of acceptance and rejection of hypotheses across Modules, and for the group considered as a whole: this pattern is shown in Table 44.14.:

Table 4.4.14: Hypotheses across sub-groups and whole group

	Whole Group	LS	ME	PS	SS	GA
H1	A	A	R	A	A	A
H2a	R	R	R	R	R	A
H2b	R	R	R	R	R	A

4.4.1. What do these findings mean?

The strong ESP construct would predict that scores on the two SAP writing tests would be highly related to each other and less highly related to scores on the GAP writing test. That is, the strong ESP construct would predict that hypothesis 1 would be rejected. On this set of data, however, hypothesis 1 was accepted for the group treated as a whole, and for all the sub-groups except Medicine. For the Medicine sub-

group there were no significant correlations between any of the tests, that is, no two writing tests shared a significant amount of common variance. Further, for this sub-group it was seen that performance on M2Q1 was inversely although weakly related to performance on SAPQ and GAPQ. Students who did well on M2Q1 did badly on both SAPQ and GAPQ.

There was no case in which a pattern of score relationships which would conform to the prediction of a strong ESP construct was found. Failure to reject hypothesis 1 supports a construct of writing proficiency as including an underlying general factor; the low correlations found in most cases, with rather small amounts of variance being accounted for, supports a construct of writing proficiency as including a number of specific factors in addition to a general factor. With this small database it was not possible to compensate for variability across tasks and individuals and conduct a factor analytic study, but the qualitative investigations which follow in Chapters 5 and 6 attempt to get closer to an understanding of what some of the specific factors might be through in-depth exploration of a number of variables.

It would seem to follow from the acceptance of hypothesis 1 that there is no reason to design and administer SAP writing tests, since scores on such tests are significantly correlated with scores on GAP writing tests, and it is much easier to prepare GAP writing tests, since one question will serve where several SAP questions are needed. However, it was also found that in every case the GAP question resulted in lower mean scores than either SAP question, which appears to be an argument in favour of SAP writing tests.

The inconsistency of these data make any conclusion difficult at this stage, and the acceptance of hypothesis 1 in all instances except Medicine must be considered in the context of the other findings.

Hypothesis 2a predicted that, since M2Q1 and SAPQ were designed to be parallel SAP tasks, they should correlate at a criterion level for equivalent forms, i.e., .80. But for this set of data it is not the case that the two SAP questions always yield equivalent scores: in fact, only in the case of the General Academic Module were we able to accept the hypothesis that M2Q1 and SAPQ were parallel forms of the same test.

We must place the rejection of hypothesis 2a alongside the acceptance of hypothesis 1 in trying to understand the interaction between the testees' knowledge and skills, and their test scores. It may be that the failure of the two SAP tests to meet the equivalence criterion accounts for the failure of the data to support a strong ESP construct in the testing of academic writing. The strong ESP construct makes no predictions about the degree of task similarity necessary to ensure equivalent forms of tests· indeed, this is an area of language testing which has to date received relatively little attention.

We can at the present moment say little about why it should be that the two SAP questions are performing differently within Modules. The ELTS operates two versions, including two versions of M2, apparently on the assumption that any questions based on this set of materials will be parallel. But this study shows clearly that it is not enough to assume task equivalence or equivalence of demands on testee knowledge and skills when two tasks are based on the same set of materials which all testees have already operated with. Neither is it possible to assume that because two writing tasks are both based in the same discipline (assuming that 'discipline' has been able to be defined prior to this, and ignoring the extreme difficulty of fulfilling this assumption) they will *de facto* be of equivalent difficulty. These assumptions require detailed investigation.

As was seen in the discussion of each Module, hypothesis 2b predicted that, although all three questions would share a significant amount of

variance, since they all have a good deal in common, i.e., they are all direct tests of writing performance of a relatively restricted, expository, kind, the amount of variance shared by the two questions designed to be 'ESP' writing tests would be greater than that shared by a putative 'ESP' writing test and a non-ESP, general academic purpose, writing test. This did not occur. Only in the case of the General Academic Module did M2Q1 and SAPQ share more variance than GAPQ and one or both of the SAPQ questions. This is a very curious exception, since by design the General Academic Module is the only Module without a specific clientele: it is taken by those testees who do not fit comfortably into any of the 'specific' Modules. Thus while each testee taking the General Academic Module has specific needs just as the testees in the other Modules do, the Module is not designed to take these into account, and we cannot usefully characterise the possible candidates in the General Academic testee pool (refer to Appendix C, 'Selection of Modular Options'). SAP questions on the General Academic Module are very difficult to distinguish from GAP questions. General Academic is the only Module where we might guess that all three questions would correlate equally highly, yet it is the only Module where the SAP ⇔ GAP difference is as predicted.

The data appear to indicate that it is not the SAP ⇔ GAP factor which causes students to perform differentially across writing tests, but again we must remember the finding that the supposedly equivalent SAP questions did not perform as equivalent in fact. These findings seem to indicate that there are several factors operating to differentiate among testees' performances across the three writing test questions. The higher mean scores for SAP than for GAP suggest that an ESP factor is playing some part, but the unpredictable correlation pattern suggests that an ESP factor alone is not a satisfactory explanation. Different questions within a Module advantage different testees, and the empirical investigation does not permit an understanding of why this might be.

4.4.2. Effect of choice of writing test on acceptance/rejection decisions

It can be seen that GAPQ always yields the lowest mean score, whether or not it has been shown to be significantly different from either or both of the SAP scores. The strong ESP construct would predict this, since according to this construct testees would be predicted, by definition as it were, to be advantaged by an ESP (SAP) test by comparison with a general (GAP) test: this is the whole argument in favour of ESP teaching/testing. On this basis, the data in this study support the strong ESP construct, although significant differences among the three writing tests means were only found for Life Sciences and Social Studies.

It would be simplistic to build an argument for SAP writing tests rather than GAP writing tests from this, however, for two reasons. First, because the amounts of variance being shared among the writing tests were typically quite small, and typically the GAP writing test shared as much or more variance with one or other of the SAP writing tests as they did with each other, we cannot be at all sure that a meaningful distinction between the constructs 'SAP' and 'GAP' is being observed. Further, we cannot be sure the degree of SAPness is constant across questions within either M2Q1 or SAPQ, since as we saw (Sections 1.2 and 2.3), we do not possess a precise model upon which the questions are based and against which they can be matched. Here we return to the problem discussed in Chapter 3, Section 4.2, where we saw that we do not have a satisfactory description of what characterises a SAP test We can only at this stage say that testees were advantaged by test x rather than test y or test z, and that test x has been claimed to be a SAP test. It may be that there are differences of absolute difficulty among the tests, and that GAPQ is simply a more difficult test: we are not in a position to claim that other variables have been held constant while only the SAPness/GAPness differs across these three writing tests.

Second, it does not follow that a test which advantages testees necessarily yields 'truer' scores than a test on which the same testees do more poorly. The ESP argument should be that an ESP test provides 'truer' scores because testees are using the same knowledge and skills that they would use in their specialist academic study; if testees do not have the requisite knowledge and skills for that specialist academic study they will presumably do poorly on the ESP test. There is therefore a prior necessity that the testees are correctly placed in a sub-group and are taking the correct writing test. We have a small amount of data to enable some consideration of this question, since the M2Q1 question for the Social Studies sub-group is also used, in ELTS Version 1, for the General Academic sub-group. Since the item was prepared for the Social Studies sub-group and then applied to the General Academic sub-group, we might predict that Social Studies testees would perform better on it than General Academic testees. This is in fact the case: the SS M2Q1 mean is 5.555 while the GA M2Q1 mean is 5.261. If we look at the ELTSOA and ELTS-M2 scores for the GA sub-group we see that they are the lowest of all sub-groups: it may be that this is the result of disadvantage caused by the inclusion in the GA source booklet of three (of the total six) texts taken from the SS source booklet. This possibility could only be explored by looking at the GA and SS scores on G1 and G2, where presumably neither group should be relatively advantaged or disadvantaged, but those data are not available to us. The implication, however, is that students in some Modules are more appropriately placed than those in others.

4.4.3. Variation across Modules

The data show that mean score levels are not constant across Modules; although the variation is not very large, the range of mean scores on M2Q1 is from 5.2 to 5.9 - effectively a range from band 5 to band 6. The variation is more extreme for both SAPQ and GAPQ. Table 4.4.15 shows the mean scores for each Module on each question for purposes of comparison:

Table 4.4.15: Writing test means compared across Modules

Module	M2Q1 $\bar{x}$	SAPQ $\bar{x}$	GAPQ $\bar{x}$
Life Sciences	5.585	6.293	4.927
Medicine	5.909	6.000	5.818
Physical Science	5.428	5.143	4.857
Social Studies	5.555	4.926	4.555
General Academic	5.261	5.826	4.782

It would seem that a testee has a better or worse chance of getting an acceptable score (where 'acceptable' refers to acceptability to university admissions officers or awarders of scholarships, who are rapidly adopting a policy of requiring 5.5/6.0 as a cutting score for applicants) depending on which Module his/her special field of study falls into. The reason for this may be that the groups differ in terms of their general language proficiency, writing proficiency, intelligence, or some other valid variable. Table 4.4.16 shows the mean ELTSOA score for each Module as the best available indicator of relative abilities of sub-groups:

Table 4.4.16: ELTSOA for sub-groups

	$\bar{x}$
Life Sciences	5.500
Medicine	6 000
Physical Science	5.714
Social Studies	5.916
General Academic	5.250

Table 4.4.16 would seem to suggest that the sub-groups are not all performing at the same level on the test as a whole, although the differences are not large. It would be dangerous to conclude anything from these differences in mean scores, but one explanation might be that testees in certain sub-groups may be stronger in some language or language-linked areas than others. Another explanation might be that testees in some sub-groups find the test (general or specific components)

more appropriate to them than others. This is not an issue which can be investigated in this study.

We would expect the M2Q1 score to be very close to the ELTSOA score, because, as explained in the First Report (February 1982, p.4), the cut-off points for the multiple-choice tests were established by matching score distributions with the score distributions on M2 for the same candidates: if the testees in this data set are performing at all similarly to the testees in the data set upon which score matching was done we are in effect seeing the writing performance influence recurring in all the objective test scores. It must be assumed that in setting score distributions for the two general multiple-choice tests the scores of all the candidates in the population on M2 were treated as a single set, since G1 and G2 do not have different standards for testees taking different Modular choices. Table 2 of the First Report (reproduced here as Appendix D) shows that score distributions were quite different from Module to Module, although the only Modular score distributions shown are for M1. Since M1 was 'normed' against M2 the score distribution for M1 must reflect the score distribution for M2 and we can consider it as representing the M2 score distribution for the testees in the standard-setting data set. Since neither mean nor raw scores are reported it is difficult to work the information provided, but of the five Modules with samples large enough for inclusion, two have the mode at band 4-4.5; one has it at band 3-3.5; one at band 5-5.5 and one at band 6-6 5 Since the population used for the First Report included testees around the world, including many who did not reach the required level to come to Britain, while the sample used in the study reported here is a truncated one, because only students who had already been accepted for study in Britain could be obtained, the range in the study is smaller and the mean is higher. It does appear, however, that the finding in this study that the sub-groups perform differently is in line with the performance of sub-groups in the operational use of the test.

If each Module and sub-group were truly independent, this would not be a problem. Score consumers could be informed of the mean score for each component of each sub-group, and use this as some sort of 'benchmark' against which to compare the scores of applicants to their institution, department, etc, or to set their own cutting scores. But the Modules/groups are not completely independent. Testees of several disciplines have difficulty in deciding which Module they should take: for example, students of Urban Design usually take the Social Studies Module, but those who have approached Urban Design through surveying and building sciences may fit better into the Technology Module. Mathematicians whose specialism is mathematical logic, set theory, etc. might be ill-suited in the Physical Science Module and opt instead for the General Academic Module. Similarly, specialists in artificial intelligence and computer languages might prefer the Social Studies Module to Physical Sciences A dietician might also prefer the Social Studies Module to the Life Sciences Module. Further, the survey of admissions officers reported by the ELTS Validation Study (Criper and Davies, 1986) found that almost all institutions followed the practice of setting a single cutting score regardless of the Module the student had taken. When such students ask for advice from the test administrator as to which Module they should take, the temptation is there for the administrator to advise them to take the 'easier' Module rather than the one closest to their future field of study.

4.4.4. Writing tests: contribution to a profile

It is also necessary to understand more about how the various writing scores interact with the ELTS score, particularly with the ELTS score without the writing sub-test. The claim for the ELTS, which is a profile test, is that the inclusion of each sub-test in a test battery increases the amount of information provided by the test. This is only true if different testees have different shapes of profiles, i.e., if it is the case that testees do not necessarily perform at the same level on

different language skills. Every time a testee displays a 'flat' profile (a profile which shows that she or he is at the same band level on every test component), a single score could be used to express her or his performance level without loss of information. It follows from this that any component of the battery that does not provide scores which are noticeably different from the scores on the rest of the components of the battery should be discarded. (It should be noted that these differences need not necessarily be significantly different: what is important to those who make admissions decisions may be differences in performances across skills which are less than statistically significant. Only for two of the five Modular sub-groups, Physical Sciences and Social Sciences, do M2Q1 scores differ more than marginally from the ELTS score without M2 (ELTS M2). One-way analysis of variance showed that for Physical Science the means were not significantly different. The means for Social Studies were significantly different at $p < .05$ (Table 4.4.17):

Table 4.4.17: ANOVA for SS writing test means with ELTS-M2

Source of variance	SS	d.f.	MS	F	p
Between groups	14.543	3	4.848	2.952	< .05
Within groups	170.816	104	1.642		

Post-hoc Scheffè comparisons ($\alpha = .05$) showed that only the ELTS-M2/GAPQ means were significantly different (Table 4.4.18):

Table 4.4.18: Post-hoc Scheffè Comparisons for Social Studies means

t_{crit}	t_{obs}		p
2.846	ELTS-M2/GAPQ	3.042	< .05

However, this way of looking at the data may not be a fair test: it may be that the combined scores on the other four sub-tests yield a mean score very similar to the writing test score, but that each separately is performing differently. The example below illustrates how this is possible:

G1	G2	M1	M3	a writing test	TOTAL
8 (+)	5 (+)	7 (+)	4	6	= 24: ÷ 4 = 6

If this were generally the case the implication would be that the writing test would be as informative as the combination of four other tests: it would be more efficient on that basis to discard the other tests and retain the writing test. The data available in this study do not permit an investigation of this question. The arguments for profile reporting, however, are not based on the relative efficiencies of single or various combinations of scores, but on a view that there are some testees who exhibit 'marked' profiles (profiles which reveal different levels of performance on different parts of the test) and that such information is of value in decision-making for both funding and admissions purposes A separate argument can of course be made for profile score reporting for diagnostic uses, especially where, as happens in the British Council context, testees are accepted with the provision that they upgrade their English by attending an English course at, or prior to commencing at, the tertiary education institution. But the relevance to this study of the question of profile reporting and of 'flat' versus 'marked' profiles is the implication for the choice of a particular writing test to form a component of the test battery rather than any other.

If a decision about the "best" test of writing were to be made on the basis of the probability of a 'marked' profile occurring, it would appear from these data that GAPQ would be the best choice. It is only in the case of the LS sub-group that GAPQ is not the writing score furthest from the ELTS-M2 score. If, however, the decision were to be made on the basis of the degreee of 'ESPness' of the test, it would appear that SAPQ would be the best choice, since of the two SAP questions it is further from the ELTS-M2 score.

Neither decision is *de facto* the right one. The fact that the multiple choice tests were 'normed' against M2 suggests that the designers of the ELTS were looking for a test battery which as a whole would generate a flat profile. We also know, however, that they were seeking to construct a test which would have at least some features of ESPness. We arereturned to questions of what should be considered a 'true' writing score, and we find that the answer remains elusive.

Clearly, the quantitative study reported here has made some of the issues of SAP versus GAP testing of writing more evident, and may have delineated them more exactly, but it has not permitted us to answer them. We are not yet at the stage where we can describe what happens in a SAP writing test which distinguishes it from a GAP writing test in terms of products, i.e., scores; still less are we able as a result of the foregoing to do this in process terms. To go further with the main research question we need qualitative study of the activities undertaken by the test maker, the test taker, and the test rater: only through such qualitative research can we understand how well ELTS M2 fulfils our expectations of it, as delineated in sub-section 1.5, and only then can we understand how well any SAP writing test might fulfil these expectations.

5. IMPLICATIONS: NEED FOR FURTHER INVESTIGATIONS

5.1. Subsidiary research question

In the process of carrying out the subsidiary study, of rater/score reliability, it had become clear that the ELTS writing test as used at that time could not guarantee adequate reliability for an operational writing test. There was no possibility of using multiple raters in the British Council context· what was urgently needed was a scoring procedure which could provide scores of adequate reliability with only one rater.

5.2. Main research question

Investigation of the scoring procedure was not only a practical necessity: the scoring procedure of a writing test is a major characteristic of the instrument, as we saw in Chapter 2, sub-section 3 2. To understand what a writing test measures, as well as to understand how well it measures it, we must look at the scoring procedure in terms of the features of writing it values or does not value. Such an investigation enables us to identify what this is a test of, we can then put this alongside a statement of what it is claimed to be a test of and consider how well it meets content and construct validity criteria. In Chapter 5, Section 1, development of a scoring procedure for use with M2 is described and the issues of score reliability and score validity are addressed.

In Chapter 5, Section 2, another major characteristic of writing tests is examined. Raters' processes and bases for judgements are studied in an attempt to understand what features distinguish the rating of a GAP writing test from the rating of a SAP writing test, and in a search for evidence that the scoring procedure used for ELTS M2 is in fact a SAP procedure.

In Chapter 6 the major question which has arisen again and again in considering these data and their interpretation, the influence of the question, is investigated.

In all essay tests, which typically are one-item or two-item tests, the actual question set is of central importance. In a SAP writing test the question will be even more critical. It is from this that the student response springs, and it is in relation to this that the scoring occurs. As we saw in Chapter 2, sub-section 3 4, work has only just begun on the description of the characteristics of essay questions. When the ELTS M2 questions were designed in 1979, even that limited amount of research was

not available to the test construction team. Even a cursory examination of the M2 questions revealed that they were not parallel: this became obvious during the attempt to construct 'parallel' questions for the second SAP essays. Questions were not parallel linguistically, cognitively or rhetorically; they were not parallel in terms of the demands they made on interpretation of the source materials or on knowledge in the specialist field. It also seemed that some questions were hardly SAP questions at all, but were in fact GAP questions. There appeared to be no way to measure the degree of discipline-specificity of questions.

We cannot expect clear results from a comparison of SAP and GAP writing tests if the test items do not themselves distinguish between SAP and GAP tasks. The one unequivocal result of the quantitative study is that the three writing tests do not maintain a stable set of relationships, let alone a stable set of relationships in the SAP ⇔ GAP dimension. A detailed investigation of the test items used in this study, within a writing test item design framework, in a search for their describable characteristics, and for some sense of their relative difficulty, may illuminate issues of test maker processes and thus the question of what constitutes an SAP writing test task and distinguishes it from a GAP writing test task.

In working with this rather small data set, a data set which is sub-divided into even smaller groupings, it was inevitable that the researcher would come very close to the essays written by the testees in the data set, and become familiar with what they said as well as how they were scored for saying it. In trying to understand why essay raters respond as they do to essay answers, again the researcher came very close to the actual content of the essays in trying to understand the rater responses. It became clear that certain essays showed characteristics which interacted with raters and their judgements in interesting ways. As the focus shifted from raters to tasks, the centrality and the chemistry of

the interaction between question and writer became clear. In particular a study which attempted to arrive at conclusions about questions and their effects without some qualitative exploration of the actual written essays began to seem limited and limiting. As the exploration of writers' responses in a search for task variables developed, it also became clear that the rater plays a part in determining what can be seen in product terms, i.e., score terms, of the question/writer interaction. Chapter 6, then, interweaves an analysis of the 6 M2Q1 questions, the 6 SAPQ questions and GAPQ with an exploration of some of the writer variables, seeking to understand what combinations of task and writer variables lead to answers which are in some sense 'SAP', and what responses 'SAP' answers generate. Thus Chapter 6 attempts to characterise the 'SAPness' or 'GAPness' of questions not only in terms of a formal task analysis, but also in part by the type of response they evoke from the writer and from the rater.

Each of these studies contributes to the attempt to say more about how ELTS M2Q1 and the other writing tests investigated fulfil the *a priori* expectations of validity resulting from the reviews in Chapters 1, 2 and 3, and the reliability expectations resulting from the reviews in Chapters 1 and 2. They also contribute to an explanation of why these expectations are not fulfilled adequately.

Following these studies, it may also be possible to draw some conclusions about how well we can expect any SAP writing test to fulfil these expectations, in comparison with any GAP writing test. The question we will seek to answer is: Given the extreme difficulty of constructing any writing test which meets these expectations, will it be more difficult, or easier, or no different, to attempt to construct specific academic purpose writing tests to meet the same expectations?

CHAPTER FIVE

PROCEDURAL VARIABLES AND READER VARIABLES

1. PROCEDURAL VARIABLES

We saw in the closing discussion of Chapter 4 that the investigation of the subsidiary research question had exposed the need for a scoring procedure for M2 which could provide adequately reliable scores with only one rater. ELTS M2 as operational in 1980 - 1985 was not fulfilling the reliability expectation. Since there are many other programmes of direct writing assessment which also have constraints of resources or contexts making it impossible for them to use more than one rater, if such a scoring procedure could be developed it could be of benefit for other contexts.

In the case of ELTS M2 a powerful additional expectation exists: construct validity. The validity claimed for ELTS in general and by extension M2 is that it is a discipline-specific (SAP) test. The accuracy of this claim was called into question by the results of the investigation of the main research question reported in Chapter 4. While working towards increased reliability for the scoring procedure, we shall also in this section pay special attention to a search for evidence of the construct validity of the procedure.

1.1. Original version

When the ELTS was first introduced, in 1980, the scoring of M2 was done with the aid of a short paragraph explaining the need to value communicative quality more than structural and surface features (see Figure 5.1.1), but the main guide for the rater was a set of performance descriptions each associated with a performance level or 'band' from 1-9 (see Figure 5.1.2), coupled with an example of performance at each level,

i.e., what is usually known as a benchmark paper. The benchmark papers are reproduced in Appendix B1.

Figure 5.1.1. First Version Scoring Advice

> *Assess with the aid of the M2 topics, Writing Assessment Scale and Writing Samples provided (see Appendix C pp 13 - 22). Use only whole, not half, bands. Judge according to the communicative quality of the writing, the effectiveness with which the arguments are presented, the logical structure of the presentation and the accuracy and appropriateness of the language used. Candidates should not be heavily penalised for making factual errors in a subject with which they may not be familiar, but answers should be relevant to the questions asked.*
>
> *Remember that it may not be possible or sensible to expound a specialist topic wholly in one's own words. (The question), however, has been worded so that it should not encourage answering by wholesale lifting from the text.* ***Wholesale lifting should be assessed as band 1 (see M2 Writing Assessment Scale). Partial lifting may contribute to an appropriate answer and should be assessed accordingly.*** *(p.5, ELTS Administrator's Manual, no date, received Autumn 1983: emphasis in the original. Appendix C appears as Appendix B1.)*

Clearly, the advice contained in this short text is unspecific and in some areas confusing, most notably the issue of the treatment of plagiarism. Many raters reported themselves unable to judge how the central administration was distinguishing between acceptable and even sensible "partial lifting" and "wholesale lifting" which called for severe penalties. Very importantly for this study, the raters are advised to pay little attention to testees' "factual errors in a subject with which they may not be familiar", placing more emphasis on relevance than on factual accuracy. This issue is discussed in detail later.

Figure 5.1.2: First Version Assessment Scale

M2 Writing Assessment Scale

BAND	BRIEF PERFORMANCE DESCRIPTION
9	Expert Writer. theme presented in a readable, intelligible, logical and interesting manner. Writes with complete accuracy and in the appropriate style. The reader is given a sense of mastery of the language and of the ability to handle the topic with complete competence.
8	Very Good Writer. theme presented clearly and logically, with accurate language forms and good style. Only very occasional inaccuracy or inappropriacy but which does not affect the communication. The reader can follow with no strain and will appreciate the argument expressed.
7	Good Writer theme presented in a well-ordered, intelligible manner with well-structured and relevant supporting detail. Generally accurate in language and appropriate in style, but occasional lapses can affect the communication on first reading. The reader has, however, the impression of a functionally efficient writer.
6	Competent Writer· theme presented fairly logically and intelligibly. Reasonably accurate use of the language system. May have inaccuracies of style and presentation but showing an adequate functional competence. Can be read with only occasional strain put on comprehension.
5	Modest Writer theme can be followed, but logical presentation may be broken and lack clarity or consistency. Several inaccuracies and style not always appropriate to presentation. May lack interest or variety, but the basic message is presented. The reader will have to strain on occasion to comprehend meaning.
4	Marginal Writer theme can be followed with effort, and closer reading reveals lack of logical structure, clarity and consistency. Inaccurate vocabulary and sentence use coupled with inadequate connectors and cohesive features. Elements of information required may be omitted, repeated or inappropriately expressed. The reader has general difficulty in working out the message, though can eventually do so.
3	Extremely Limited Writer. elements of the information required are provided, but the presentation lacks any coherence. Uses over-simple sentence structure and impoverished vocabulary with continual errors and inappropriateness. Below level of functional competence though the reader may work out the general message.
2	Intermittent Writer: elements of the information required not provided, although a general meaning comes through intermittently. Either copies or produces strings of words. No real communication with the reader having constant problems in making out any message.
1	Non-Writer cannot write the language. OR: cannot be adequately assessed either because answers have been lifted 'en bloc' from the Source Booklet, or because a clearly irrelevant stock answer has been produced.
0	No questions have been attempted.

(p. 19, ELTS Administrators' Manual, op cit)

The Assessment Scale is an attempt to establish criterion performance descriptions for each point on the scale. Each description is also summarised by a performance label (Very Good Writer, etc). As we saw in Chapter 4, the assumption was that a direct writing test coupled with these performance descriptions was *de facto* reliable, that because the test was criterion-referenced rather than norm-referenced, questions of reliability were superfluous. So great was the confidence in the intrinsic meaningfulness and accuracy of the scores that the three objective tests were normed against them.

But there are several problems with the Assessment Scale. First, each performance description in the Assessment Scale is fairly lengthy and refers to several features and qualities of writing. The references to various features and qualities of writing do not appear in the same sequence in every description in the Scale. It is not always possible on the basis of the description in the Scale to distinguish performance on one aspect of writing on two adjacent points on the Scale (e.g., band 6 "Can be read with only occasional strain put on comprehension " and band 5 "The reader will have to strain on occasion to comprehend meaning."). Most significantly, there is an assumption inherent in the Assessment Scale that every writer has a uniform level of writing profiency, that is, that because a writer can be characterised by, for example, "The reader is given a sense of complete mastery of the language..." it is necessarily also true that the reader has a sense of"...the ability to handle the topic with complete confidence." (band 9). A single scale implies a unidimensional view of writing proficiency, and necessitates the treatment of each essay as existing at a single level. But reports from raters indicated that there were occasions where a rater had difficulty rating an answer because she could not see one uniform level in it: it appeared that certain answers did not demonstrate unidimensional writing proficiency. This problem resulted in widely differing assessments of the same answer for certain non-unidimensional answers.

In addition, the labels 'Competent Writer', 'Marginal Writer', etc. are difficult to interpret: a study of 20 raters new to M2, conducted informally by this researcher, showed that only 14 were able to correctly match the labels with the descriptions and number the sets in the correct order 1-9. The labels 'Marginal', Extremely Limited', and 'Intermittent' caused the most problems. Further, it appeared that the labels tended to discourage raters from looking closely at the full performance descriptions by appearing to encapsulate all the information in the description.

While this researcher's discussions with the raters who did the rating in this study, coupled with her experiences training raters for M2 scoring as part of the data collection for the EEVP, had emphasised the problems discussed above, the British Council was experiencing similar problems in reports from raters in the field. It became clear that raters, most of whom were British Council ELT officers working in centres outside Britain and often in isolation from other M2 raters, needed firmer and more consistent guidance in rating M2. A full Assessment Guide for M2 was clearly necessary, not only for future research into the test, but more urgently for operational use.

1.2. Second version

The second version of the scoring procedure took the form of an extensive combined training manual and rating guide. This Assessment Guide, as it became called, took account of classical and more recent studies in the direct assessment of writing, as these were surveyed in Chapter 4, and also worked within the constraints imposed by the operational and administrative needs of the British Council/University of Cambridge Local Examinations Syndicate.

1.2.1. Making the criteria/traits explicit

The major development in the Assessment Guide was to take the criteria which had been implicit in the original general explanation of what should be valued in M2 answers and make them explicit. Study of the scoring advice in the first version (Figure 5.1.1) suggested that the test constructors had intended to identify four criteria, or traits of writing proficiency, to which raters should respond: communicative quality, effectiveness of argumentation, logical structure (organisation), linguistic accuracy and appropriacy. Each of these criteria was extensively characterised as shown in Figure 5.1.3 (note that all Figures in Section 1.2. are taken from ELTS: Assessment Guide for M2 Writing, March 1985 and March 1986):

Figure 5.1.3: Second Version Criteria

In scoring M2 writing, we are interested not only in grammar and vocabulary, but also in the higher communicative skills such as logical structure and presentation of ideas. The criteria to be looked for in any answer are:

(a) communicative quality
This is the most difficult criterion to describe because it is not separable from the script, but comes through, to a greater or lesser extent, in the overall effect of the answer. It is the most general, the most global, impression you get of the ability of the candidate to make the message clear to the reader. The other criteria play an inevitable part in this, but in judging this criterion you should not look specifically at any of the other criteria.

(b) organisation
This is the extent to which the logical structure chosen by the candidate to carry the message enables you to follow a thread through the answer, and to see which points the candidate thinks are important and which less so. Organisation is helped by paragraphing conventions, by adequate and flexible sentence length, by introducing new ideas in relation to previously explained ideas, by stating the topic

clearly and early, and by using linguistic coherence features (e.g. First ... Second ... Third ...; While ... On the other hand ...)

(c) *argumentation*
The effectiveness of the arguments the candidate uses to respond to the question is extremely important in establishing and developing a meaningful and interesting answer. Although 'originality' is restricted by the kinds of questions asked, each candidate should be able to bring in some new information or personal reaction, and to relate to the topic of the question in an individual way. If the arguments the candidate gives you are confused and contradictory this will lower your impression of him/her on this criterion, because it will make any message more difficult to understand.

(d) *linguistic accuracy and appropriacy*
The accuracy of syntax and spelling and the use of suitable vocabulary and punctuation are important to satisfactory communication in writing. In assessing the significance of grammatical and other linguistic errors more attention should be paid to the damage these do to communication than to their frequency. Linguistic appropriacy refers to the effectiveness of the grammatical and lexical choices the candidate has made. It should also be judged according to the way it contributes to effective communication rather than by reference to theoretical concepts such as vocabulary difficulty level.

Although all the criteria are equally important to the judgement you make, it can easily be seen that because communicative quality is a more global criterion, and because all the other criteria contribute to it, it will play the greatest part in influencing your decision.
(p. 2-3)

1.2.2. Establishing a scoring procedure

Accepting the unsuitability of a true holistic scoring procedure in the British Council context, where the multiple marking necessary to ensure adequate reliability was not a possibility, an alternative approach which could improve the reliability of M2 was needed. The procedure that was

developed required the same rater to read several times, looking at different features each time, thus attempting a mutiple response from a single rater. The exact procedure is given in Figure 5 1.4:

Figure 5.1.4.: Second Version Global Method

(i) *Read the script quickly looking specifically at its communicative quality: what is the message? is it successful? is it reasonable? Spend NOT MORE THAN 2 minutes. Immediately decide to yourself a range of bands for the paper - use 3 bands (e.g., it's good, but not perfect - somewhere from 8 to 6; or, it's rather ordinary, not too bad but not outstanding - between 7 and 5).*

(ii) *Now read the script again, also in not more than 2 minutes. Pay particular attention to organisation and argumentation Decide whether to knock off the highest or the lowest of the 3 bands you assigned (e.g., well-organised, good arguments - no basis for a 6, so between 8 and 7; or, arguments aren't bad, but it's badly organised - not worth a 7, so between 6 and 5).*

(iii) *Read the script a third time, again in under 2 minutes. Pay attention to linguistic appropriacy and linguistic accuracy. Decide which of the two bands is your final assessment (e.g., one or two linguistic inaccuracies of complex grammar, but very appropriate - this is a band 8; or, quite a few grammar mistakes though they don't interfere with communication, but inappropriate vocabulary which interferes with the message - band 5). (p.4)*

The procedure of first deciding on a three-band range and then narrowing it parallels the procedure used for assessing M3 (Oral Interview). It guides the rater to focus attention at a certain text level on each reading occasion, and to use the decision at each stage as input for the decision at the next stage.

1.2.3. Toward a multiple trait procedure

The decision had been made, by the British Council/University of Cambridge Local Examinations Syndicate, not to include a revision of the original assessment scale in the brief for the development of the second version scoring procedure. Therefore, in an attempt to provide an alternative solution to the problems with the Assessment Scale described in sub-section 1.1, the Assessment Guide incorporated a second scoring procedure to be applied when the rater feels that the writer demonstrates greater proficieny on some criteria than on others. This is described as a 'marked' profile, a term which was introduced into ELTS to describe variations in proficiency across skills, but which experience has shown to be observable across the dimensions of writing skill. In this procedure, known as the 'Profile Method', raters first apply the 'Global Method': they then apply this procedure, following the instructions given in Figure 5 1 5:

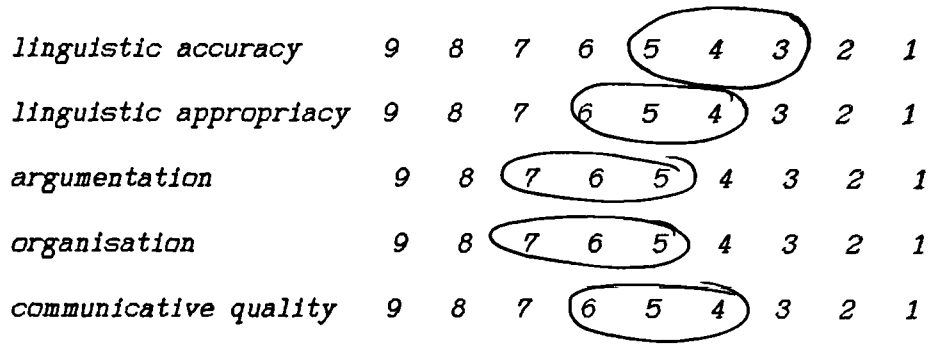
Figure 5 1.5 : Profile Method

1. Skim the script again as many times as you need to, but very quickly each time, until you have been able to circle a 3-band range on each line of the profile grid below:

linguistic accuracy	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*
linguistic appropriacy	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*
argumentation	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*
organisation	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*
communicative quality	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*

2. Now there are two ways to reach a final band, and you may choose whichever seems appropriate:

(i) if you see a clear pattern on the grid which enables you to make up your mind on the best band, do so. For instance, if you found you had this pattern:



you would see that weak grammar had lowered an assessment which was generally higher, and that in every criterion band 5 was one of those circled. You could decide to assign a final band 5.

(ii) alternatively, you may prefer to take the mid-band in each case, total and average them. In the example above this yields

4 + 5 + 6 + 6 + 5 = 26 ÷ 5 = 5.2 rounded to band 5.

The two systems do not always yield the same result, especially in extremely marked profiles. In these very rare cases the second method is recommended.

1.2.4. Dealing with problems

Raters had raised a number of key problems, and the piloting had both shown that the problems were real and suggested some guidelines to help raters with some of them. These problems were: length, irrelevance, factual errors, and plagiarism.

1.2 4.1. Length

During the piloting of the scoring procedure with a sample set of essays, there was much discussion of the minimum acceptable length for an answer. The original instruction in the rubric had been to write "15 to 20 lines", but the number of words per line was very varied, and some students were writing only 40+ words as an answer. The minimum length

was set at 60 words, below which raters are advised that they are probably faced with an insufficient sample upon which to base a decision.

1.2.4.2. Irrelevance

Raters are told that irrelevance occurs for two reasons. The first is that the testee did not understand the question, or the input text; or she has learned a 'stock answer' which does not relate to the question but is the result of some second- or third-hand information about what to expect (this problem increases as the test questions are seen and partially recalled by more and more candidates, and suggests the need for more frequent introduction of alternate questions). Inability to understand the question or the input text are discussed as task variables in Chapter 6.

The second reason is that the candidate "objects to or is unable to answer the question" (Guide, p 7). In a truly SAP writing test, we would not expect irrelevance due to inability to answer the question to occur; when it does occur it may be considered an indication of a mismatch between the writer and the particular test item (essay question). If it occurs frequently, it will call into question the design of the test Irrelevance resulting from a candidate's objection to or 'challenge' of the test is dealt with in Chapter 6 as a task variable.

We shall see in the next section that, even after raters had been offered advice in the Assessment Guide about the handling of irrelevance, it remained a major problem.

1.2.4.3. Factual errors

The question of factual errors, problematic in the original scoring advice, was handled as shown in Figure 5.1.6:

Figure 5.1.6: Factual Errors

> *Where a candidate makes factual errors which indicate failure to comprehend the input text, and where these errors reduce the effectiveness of the argument, the candidate's score will be affected. The extent to which such errors result in a lower score depends on how important the point was to the overall answer.*
>
> *Where a candidate makes factual errors on information which he/she has drawn in from outside the text, i.e., from their own knowledge, these errors should not be penalised. In these cases you need to try to suspend your own outside knowledge and accept the information as legitimate for the immediate task. This is, of course, sometimes very hard to do, but the attempt should be consciously made. (p. 7)*

Clearly, there are still problems here, and these problems relate very directly to the central question of this study: the argument for a SAP writing test rather than a GAP writing test is that in a SAP writing test the writer can draw on her own expertise in the specialist area. The handling of factual errors shows the inescapable tension between the desire not to disadvantage the testee who is not familiar with the specific content selection from the much wider content area, and the desire to create a discipline-specific 'feel' to the essay questions. Here the suggestion that the language and content can be separated is overtly made, and in the separation the language is to be valued without regard to the content. The advice to "accept the information as legitimate for the immediate task" relates directly to questions of relevance, and thus leads the rater not to a solution but to another set of problems.

The solutions to these tensions and problems, if there are any, lie not in manipulations of the scoring procedure but in careful attention to the interaction between task design and scoring criteria which are valid for the assessment of the specific task, that is, criteria which are salient for the assessment of the kind of writing performance the task is intended to elicit: we shall return to them in Chapter 6.

1.2.4.4. Plagiarism

Plagiarism, which had also been handled in a problematic way in the first version (wholesale lifting vs. partial lifting), was dealt with at some length in the Assessment Guide (Figure 5.1.7).

Figure 5.1.7: Plagiarism

> *Plagiarism refers to 'lifting' or 'stealing' the actual language of the input text. It is quite permissible to use the ideas and information of the input text, as far as these are relevant, but in all cases the language should be paraphrased and reordered so that it is expressed within the candidate's organisational and argumentative structure, and so that it is compatible with his/her overall tone, style and linguistic level.*
>
> *It is difficult to set absolute rules about what constitutes plagiarism. In the first place, you need to be extremely familiar with the input text before beginning to assess scripts. This familiarity will aid the recognition of plagiarism. Plagiarism is also usually easy to recognise because of the linguistic and stylistic mismatch of this with the candidate's own writing. Once plagiarism is suspected, you should compare the answer to the input text to discover how much of the candidate's answer is dependent on direct lifting. This can be quantified by judging what percentage of the answer is truly in the candidate's own words, and also by judging how many of the points made by the candidate are taken word-for-word from the input text. Where all or virtually all the candidate's answer is lifted directly from the text, it should be assessed as band 1. Where the candidate has relied on the input text for all or most of the arguments but has added some material, the penalty should be severe (i.e., lower the overall band by 2-3 bands from what it would have been worth if original). Where a large part of the language at clause /phrase level is lifted, but used to support the candidate's own arguments and within his/her own organisational structure, the penalty may be less severe (i.e., lower the overall band by 1-2 bands depending on the quantity of lifting). If a candidate has plagiarised occasional clauses or phrases (as a rough guide, less than 10% of the total answer) but built these into her/his answer effectively, no penalty should be applied. (p. 8)*

We shall see in Section 2 that raters made only minimal mention of plagiarism: it is not possible to know whether this means that the advice given here allowed raters to solve their problems with plagiarism to their own satisfaction, or whether such problems did not arise in the limited data set raters worked with.

1.2.5. Piloting the second version

When the new scoring procedures had been worked out, and procedures for dealing with problems set up, the procedure was piloted. Four raters with varying amounts of M2 rating experience worked with the new procedure on a set of 23 M2Q1 answers. They did this first individually, and then were gathered together to discuss their scores.

To do this, they were allowed two or three minutes to refamiliarise themselves with each script, without being aware of the score they had given it on the first read, and were then asked to comment briefly on the answer and give their score. After all scores had been given discussion was invited; sometimes specific questions were addressed to specific raters. All of the score-giving and discussion was tape-recorded.

The researcher took away the scores given by all raters on each occasion and the audio tapes, and used the information to: (a) select the answers for inclusion in the Assessment Guide, either in the 'criterion set' of 10 answers which would be used for initial training, or in the 'sample set' of 10, which would be used by raters needing to refamiliarise themselves after a period without doing any rating; (b) prepare a summary of the discussion among the raters for each answer chosen for inclusion in the criterion set, to enable raters training by themselves to have some idea of how other raters responded to the criterion set, and why, as well as knowing what scores were decided upon. Figure 5.1.8 gives an example of such a summary, for answer 9 in the criterion set: the answer itself is included in Appendix E.

Figure 5.1.8.: Example Using Global Method

> *The marking team felt that this script began at a functionally competent level of communicative quality but deteriorated particularly in the last 'paragraph'. This seems to be a writer who cannot yet sustain a written discussion in English. There are signs of competent organisation, and the first paragraph is adequately argued. An initial 5-4-3 was assigned, which was reduced to 4-3 on the basis of the problems in all the areas of communicative quality, organisation, and argument in the second half. This was reduced to a final band 3 because of the level of linguistic inaccuracy and inappropriacy. (p. 10)*

The criterion set of answers included one which the marking team had felt they needed to apply the 'Profile Method' to, answer 6. The summary of their discussion is shown in Figure 5.1.9, and the answer appears in Appendix E:

Figure 5.1.9.: Example Using Profile Method

> *The markers used the Profile Method for this script because they found it difficult to decide between bands 3 and 4 on the Global Method. In the discussion it became apparent that although the script gives the impression of being well organised and strongly argued, this is not the case The arguments basically repeat the same point; they are also based on unfounded assumptions and are not well expressed. Markers thought the paper was very limited linguistically, and particularly weak on connections between sentences. An examination of the Profile Grid for the script supported markers' impression that the paper was, unusually, low on communicative quality compared to other features:*

linguistic accuracy	9	8	7	6	5	4	3	2	1
linguistic appropriacy	9	8	7	6	5	4	3	2	1
argumentation	9	8	7	6	5	4	3	2	1
organisation	9	8	7	6	5	4	3	2	1
communicative quality	9	8	7	6	5	4	3	2	1

Based on the Profile, the final band awarded was 4.

As a result of the piloting, a number of minor changes to the language and sequence were made, but the main outcome of the piloting was the selection and sequencing of the answers for the criterion set and the sample set. Because a number of answers were discarded, two additional answers had to be selected and scored against the new procedure prior to their inclusion in one set or the other.

1.3. Third version

When the Assessment Guide had been in operation for a year, the researcher was invited to prepare a similar Guide for the second question of M2 (not dealt with in this study). While the sense from the British Council and UCLES based on operational use of the Assessment Guide was that it had improved matters considerably, the researcher felt that there were three main reasons to further develop the scoring procedure for M2Q1 at the same time.

First, the intention had been that the Profile Method would only be used with problem essays, after an initial application of the Global Method. It became clear, however, that some raters began to apply the Profile Method to every paper. This meant that instead of, as had been intended, looking first at communication quality in the Global Method and then

moving through organisation and argumentation to linguistic appropriacy and accuracy (i.e., macro to micro features), these raters began with linguistic features and moved in the opposite direction This resulted in more emphasis being placed on linguistic features than had been intended by the test design. We may speculate as to the reasons for this preferred use of the Profile Method: it may be because there is no such thing (or, at least, that some raters perceive no such thing) as a 'flat' profile, i.e., a writer whose proficiency is the same on every aspect of the writing process, or it may be that the Profile Method artificially creates multiple samples, permitting an objectivisation of what is for some raters an uncomfortably subjective process.

Second, the criterion 'linguistic accuracy and appropriacy' had been separated out in the profile grid into 'linguistic accuracy' and 'linguistic appropriacy': this was because raters had reported that they sometimes found that a writer used very accurate linguistic forms which were not appropriate to the task, usually for reasons of register, or that a writer might show a strong sense of appropriacy of language but not be fully in control of the accuracy dimension. However, this separation meant that linguistic characteristics were weighted heavily in the Profile Method, while in the Global Method, although the modified holistic procedure did not specify any weightings, the intention was that linguistic characteristics would be less heavily weighted than the criteria described first.

But the most critical reason for further development was the need for revision of the assessment scale. We saw these problems in sub-section 1.1, and the Profile Method was intended to reduce them. The criteria were not fully or consistently articulated in the original assessment scale: the test designers had themselves been searching for a sense of what the criteria were or should be, and this was only known as a consequence of the operationalisation of the test. Revision would permit the assessment scale to be brought into line with the rest of the scoring

procedure, providing a clear and consistent treatment of the same criteria. Further, in the original assessment scale the nine levels of performance were not clearly or consistently differentiated: raters had reported that they found difficulty differentiating between bands 6 and 5 in particular. Revision would permit the clear differentiation of the nine levels on each of the criteria.

1.3.1. Redefining the criteria/traits

Following discussions with the British Council/University of Cambridge Local Examinations Syndicate it was agreed that there were sufficient indications that raters found occasion to apply two separate linguistic criteria, accuracy and appropriacy and that these two should be treated as separate in the revision of the assessment scale. When these two were separated out the criteria were stated as shown in Figure 5.1.10. (Figure 5.1.10 and all other extracts in this section are taken from the manuscript of ELTS: Assessment Guide for M2 Writing, to appear in February 1987):

Figure 5.1.10.: Second Version Linguistic Criteria

(d) *linguistic appropriacy*

Linguistic accuracy refers to the effectiveness of the grammatical and lexical choices the candidate has made, in relation to the demands of the specific question being answered. Appropriacy should be judged by the way these choices contribute to effective communication rather than by reference to theoretical concepts such as vocabulary difficulty level. Sometimes a candidate can be seen to be using grammatical patterns which are acceptable, but not the most efficient ones to convey the intended message. The same is true of lexical choices. This may be an indication of limitations on the candidate's control of the appropriate grammar and lexis. Where the question demands reference to particular information in the input text, the candidate may find it necessary to use specialist lexis found in the text. Such choices, where there

is a precise term for a concept, are fully appropriate (see 'Plagiarism', below).

(e) *linguistic accuracy*
Grammatical choices should not only be appropriate, but also accurate in terms of, for example, subject/verb agreement, choice of tenses, clause/phrase structure, correct position of adverbs, etc. The linguistic accuracy criterion also includes spelling and punctuation. In assessing the significance of errors of grammar, spelling and punctuation more attention should be paid to the damage these do to communication than to their frequency.
(ms.p. 29)

1.3.2 Developing the new assessment scales

It was decided that two assessment scales would be needed. one, to parallel the original assessment scale, for use in the Global scoring method; the other, to be a development from the profile grid used in the second version for the Profile Method. It seemed most sensible to develop the profile scale first, and later to put this together to form the global scale.

The profile scale was developed by writing a performance description for each trait at each level, attempting to make each performance description distinguishable from the adjacent band levels on the same criterion. The five components of the profile scale were revised four times after review by the British Council/University of Cambridge Local Examinations Syndicate language testing specialists or trialling with specimen answers. The final version of the profile scale is shown in Figure 5.1.11.

The global scale was constructed directly from the profile scale very simply, by joining together the five descriptions of the five traits at each level. Sometimes no further work was necessary; sometimes some

minor stylistic changes were needed. The final version of the global scale is shown in Figure 5.1.12.

When the two scales were completed, an informal trial with 20 inexperienced raters showed that all 20 were able to correctly sequence the performance descriptors for each of the traits and then combine these to re-form the global scale.

Figure 5.1.11.: Profile Scale

DRAFT

PROFILE SCALE

COMMUNICATIVE QUALITY	ORGANISATION	ARGUMENTATION	LINGUISTIC ACCURACY	LINGUISTIC APPROPRIACY
9. The writing displays an ability to communicate in a way which gives the reader full satisfaction.	The writing displays a completely logical organisational structure which enables the message to be followed effortlessly.	Relevant arguments are presented in an interesting way, with main ideas prominently and clearly stated, with completely effective supporting material; arguments are effectively related to the writer's experience or views.	The reader sees no errors of vocabulary, spelling, punctuation or grammar.	There is an ability to manipulate the linguistic systems with complete appropriacy.
8. The writing displays an ability to communicate without causing the reader any difficulties.	The writing displays a logical organisational structure which enables the message to be followed easily.	Relevant arguments are presented in an interesting way, with main ideas highlighted, effective supporting material and they are well related to the writer's own experience or views.	The reader sees no significant errors of vocabulary, spelling, punctuation or grammar.	There is an ability to manipulate the linguistic systems appropriately.
7. The writing displays an ability to communicate with few difficulties for the reader.	The writing displays good organisational structure which enables the message to be followed without much effort.	Arguments are well presented with relevant supporting material and an attempt to relate them to the writer's experience or views.	The reader is aware of but not troubled by occasional minor errors of vocabulary, spelling, punctuation or grammar.	There are minor limitations to the ability to manipulate the linguistic systems appropriately which do not intrude on the reader.
6. The writing displays an ability to communicate although there is occasional strain for the reader.	The writing is organised well enough for the message to be followed throughout.	Arguments are presented but it may be difficult for the reader to distinguish main ideas from supporting material; main ideas may not be supported; their relevance may be dubious; arguments may not be related to the writer's experience or views.	The reader is aware of errors of vocabulary, spelling, punctuation or grammar, but these intrude only occasionally.	There is limited ability to manipulate the linguistic systems appropriately, but this intrudes only occasionally.
5. The writing displays an ability to communicate although there is often strain for the reader.	The writing is organised well enough for the message to be followed most of the time.	Arguments are presented but may lack relevance, clarity, consistency or support; they may not be related to the writer's experience or views.	The reader is aware of errors of vocabulary, spelling, punctuation or grammar which intrude frequently.	There is limited ability to manipulate the linguistic systems appropriately which intrudes frequently.
4. The writing displays a limited ability to communicate which puts strain on the reader throughout.	The writing lacks a clear organisational structure and the message is difficult to follow.	Arguments are inadequately presented and supported; they may be irrelevant; if the writer's experience or views are presented their relevance may be difficult to see.	The reader finds the control of vocabulary, spelling, punctuation and grammar inadequate.	There is inability to manipulate the linguistic systems appropriately, which causes severe strain for the reader.
3. The writing does not display an ability to communicate although meaning comes through spasmodically.	The writing has no discernible organisational structure and a message cannot be followed.	Some elements of information are present but the reader is not provided with an argument, or the argument is mainly irrelevant.	The reader is primarily aware of gross inadequacies of vocabulary, spelling, punctuation and grammar.	There is little or no sense of linguistic appropriacy, although there is evidence of sentence structure.
2. The writing displays no ability to communicate.	No organisational structure or message is recognisable.	A meaning comes through occasionally but it is not relevant.	The reader sees no evidence of control of vocabulary, spelling, punctuation or grammar.	There is no sense of linguistic appropriacy.
1. A true non-writer who has not produced any assessable strings of English writing. An answer which is wholly or almost wholly copied from the input text or task is in this category.				
0. Should only be used where a candidate did not attend or attempt this part of the test in any way.	Liz Hamp-Lyons for the British Council ©British Council 1986			

Figure 5.1.12.: Global Scale

DRAFT

GLOBAL SCALE

9	The writing displays an ability to communicate in a way which gives the reader full satisfartion It displays a completely logical organisational structure which enables the message to be followed effortlessly Relevant arguments are presented in an interesting way, with main ideas prominently and clearly stated, with completely effective supporting material, arguments are effectively related to the writer's experience or views There are no errors of vocabulary, spelling, punctuation or grammar and the writing shows an ability to manipulate the linguistic systems with complete appropriacy
8	The writing displays an ability to communicate without causing the reader any difficulties It displays a logical organisational structure which enables the message to be followed easily Relevant arguments are presented in an interesting way, with main ideas highlighted, effective supporting material and they are well related to the writer's own experience or views There are no significant errors of vocabulary, spelling, punctuation or grammar and the writing reveals an ability to manipulate the linguistic systems appropriately
7	The writing displays an ability to communicate with few difficulties for the reader It displays good organisational structure which enables the message to be followed without much effort Arguments are well presented with relevant supporting material and an attempt to relate them to the writer's experience or views The reader is aware of but not troubled by occasional minor errors of vocabulary, spelling, punctuation or grammar, and/or some limitations to the writer's ability to manipulate the linguistic systems appropriately
6	The writing displays an ability to communicate although there is occasional strain for the reader It is organised well enough for the message to be followed throughout Arguments are presented but it may be difficult for the reader to distinguish main ideas from supporting material, main ideas may not be supported, their relevance may be dubious, arguments may not be related to the writer's experience or views The reader is aware of errors of vocabulary, spelling, punctuation or grammar, and/or limited ability to manipulate the linguistic systems appropriately, but these intrude only occasionally
5	The writing displays an ability to communicate although there is often strain for the reader It is organised well enough for the message to be followed most of the time Arguments are presented but may lack relevance, clarity, consistency or support, they may not be related to the writer's experience or views The reader is aware of errors of vocabulary, spelling, punctuation or grammar which intrude frequently, and of limited ability to manipulate the linguistic systems appropriately
4	The writing displays a limited ability to communicate which puts strain on the reader throughout It lacks a clear organisational structure and the message is difficult to follow Arguments are inadequately presented and supported, they may be irrelevant, if the writer's experience or views are presented their relevance may be difficult to see The control of vocabulary, spelling, punctuation and grammar is inadequate, and the writer displays inability to manipulate the linguistic systems appropriately, causing severe strain for the reader
3	The writing does not display an ability to communicate although meaning comes through spasmodically The reader cannot find any organisational structure and cannot follow a message Some elements of information are present but the reader is not provided with an argument, or the argument is mainly irrelevant The reader is primarily aware of gross inadequacies of vocabulary, spelling, punctuation and grammar, the writer seems to have no sense of linguistic appropriacy, although there is evidence of sentence structure
2	The writing displays no ability to communicate No organisational structure or message is recognisable A meaning comes through occasionally but it is not relevant There is no evidence of control of vocabulary, spelling, punctuation or grammar, and no sense of linguistic appropriacy
1	A true non-writer who has not produced any assessable strings of English writing An answer which is wholly or almost wholly copied from the input text or task is in this category
0	Should only be used where a candidate did not attend or attempt this part of the test in any way (i e did not submit an answer paper with his/her name and candidate number written on)

Liz Hamp-Lyons for the British Council

1.3.3. 'Global' scoring procedure

The scoring procedure using the global scale is the same as that in the second version. It was predicted, however, that more reliable scores would result when the procedure was used with an assessment scale which directly matches the statement of criteria in the training information for the procedure.

1.3.4. 'Profile' scoring procedure

Although the use of the profile scoring procedure is recommended in the same circumstances as those in which the Profile Method was recommended in the second version, the scoring procedure using the profile scale differs from that applied in the second version when using the Profile Method. The rater has both the profile scale and a revised profile grid to work with. The revised profile grid presents the traits in reverse order compared to the profile grid developed for the second version (Figure 5.1.13):

Figure 5.1.13.: Revised Profile Grid

communicative quality	9	8	7	6	5	4	3	2	1
organisation	9	8	7	6	5	4	3	2	1
argumentation	9	8	7	6	5	4	3	2	1
linguistic appropriacy	9	8	7	6	5	4	3	2	1
linguistic accuracy	9	8	7	6	5	4	3	2	1

A second difference is that the rater is asked to choose a single band to describe performance on each criterion, not the three-band range previously used. The imprecision of the former procedure added to the difficulties of score aggregating, and trialling showed that raters found the profile descriptors sufficiently precise that they felt able to work

with a single band. Otherwise, the new Profile Scale presents the same problems of score aggregating as the Profile Method did.

1.3.5. Score aggregating

There is no satisfactory mathematical formula which can be applied in aggregating the scores on the five traits when using the new Profile Method; combining scores on organisation and linguistic accuracy and calling the answer writing proficiency is much like adding two apples and three pears and calling the result a lemon. Nevertheless, it has to be done, since clearly those responsible for absolute acceptance/rejection decisions for university places or for scholarships must have a single number to use. Whatever ethics or aesthetics may desire, this is the practical reality. It must be the test developers' responsibility to advise the score consumers of their best estimate of the candidate's writing proficiency, treating as unidimensional that which experience has shown is not unidimensional. The way in which the separate scores are aggregated must reflect the belief of the test developers about what is important in writing performance for the specific context, and in what proportions compared to other dimensions entering the same equation. There is no single 'right answer'. The answer which was arrived at for the particular ELTS M2 context was to weight communicative quality twice and all the other criteria once. However, no-one involved believes this is an insignificant decision, and it is one which will be monitored continually as the new procedure becomes operational.

1.3.6. Piloting the third version

The third version was piloted in the same way as the second version but using different raters, two of whom were very highly experienced with the second version, and two who had only limited rating experience. Piloting occurred on two occasions, first to trial the revisions to the procedures, and second to apply the revised procedures to a rescoring of the answers

used in the first Assessment Guide and some additional answers. Once again, all sessions were audio-taped. The profile scale was also sent to these raters for their comments during the preparation stages. The instructions for the treatment of 'stock answers' was altered, allowing relevant stock answers to be scored in the usual way. this was because it was agreed, afetr lengthy discussion, that when no actual evidence was available to show that the candidate had behaved dishonestly, the candidate could not be penalised. Apart from some linguistic improvements in the performance descriptors (mentioned above), and some rethinking of the applications of the lowest levels of the scales, piloting did not result in further changes.

1.4. Applications of the multiple trait procedure

The profile scale and what is called in the M2 Assessment Guides the Profile Method arose from the conviction of raters that it was not always possible to state with any confidence the writing proficiency level of a writer in holistic terms, i.e., as a single trait. Work with raters during development of the second and third versions suggests that about 1 in 10 answers are difficult to score using the global scale only. In these answers the writer appears to demonstrate multiple writing proficiencies, and the rater feels the need to acknowledge and respond to these multiple proficiencies separately. When the scores given by the rater on each of the traits are combined, the result should not be thought of as representing 'overall' writing proficiency: an aggregated score may fall at a band level which does not represent the writer's actual performance on any of the traits measured separately. An example of how this can occur is given in Figure 5.1.14:

Figure 5.1.14.: Aggregating a Marked Profile

communicative quality	9	8	(7)	6	5	4	3	2	1
organisation	9	8	7	6	(5)	4	3	2	1
argumentation	9	8	7	6	5	(4)	3	2	1
linguistic appropriacy	9	(8)	7	6	5	4	3	2	1
linguistic accuracy	9	(8)	7	6	5	4	3	2	1

AGGREGATE SCORE *(7 x 2) + 5 + 4 + 8 + 8 = 39 - 6 = 6.5*

The evidence suggests, as will be seen in sub-section 1.5 of this Chapter, that the use of the multiple trait procedure followed by score combining yields more reliable scores in those cases where multiple writing proficiency is demonstrated in the answer. If this was the only application of the procedure it would be valuable enough to justify its use.

But the multiple trait procedure also has another, potentially equally important, application A testing system such as the ELTS is predicated on the belief that by administering tests of different skills, using different methods, and by reporting scores on each of these tests, not simply more but also better information is obtained about candidates, and as a result better decisions are made. As we saw in the example above (Figure 5.1.14), when scores are aggregated information is lost. the aggregate does not precisely represent the rater's judgement. Scores generated by the multiple trait procedure are most valid when they are reported and interpreted separately.

The reporting of separate scores on the multiple trait procedure is an extension of the profiling construct, which is at the heart of the ELTS, to a further level, from across skills to within what is traditionally thought of as a single skill. Such an extension has practical applications. Applicants for British Council funding are usually applying

for specific courses of study, and all the information yielded by the test scores can be used when considering whether to accept or reject the applicant, or to give a qualified acceptance with some mandatory English language study prior to commencement of the main course of study. Further, when such a qualified acceptance takes the student into either a pre-sessional course at a quality institution in Britain or language courses in a British Council teaching centre outside Britain, the finely-grained information from the skills profile, and equally from the multiple trait procedure on M2, can be applied for diagnostic purposes. The multiple trait procedure allows language tutors to know, firstly, whether or not this is a student with a 'marked' writing profile, i.e., whether the student is likely to have specialised needs or to be able to be integrated into a writing class with others at a similar overall writing band level.

Then, if a 'marked' profile indicates special needs, language tutors can look at that profile and interpret it in terms of how they can work with the student as an individual within the curriculum of their institution. A student with a weak score on linguistic accuracy could be placed into additional courses in grammar and vocabulary study, a student with a weak linguistic appropriacy score could take extra reading courses or conversation courses to gain more exposure to linguistic structures and lexis in use; one with poor organisation skill might be helped by a study skills or introductory research methods course. The multiple trait procedure, then, leads to improved reliability, validity and diagnosis.

1.5. Pilot validation of scoring procedures

When the development of the third version of the Assesment Guide was complete, a small preliminary study was conducted to compare the various scoring procedures developed and used for scoring M2Q1 over its life to that point.

1.5.1. Design of the study

Twelve inexperienced raters worked in four teams of three, each team using a different scoring procedure. All teams received the same orientation to ELTS and to M2, then each team received a brief orientation to the scoring procedure they were to use, together with copies of whatever printed training material existed for the procedure they were using. The raters were chosen mainly for their availability and willingness, but also as being suitable candidates for positions in British Council DTO's (Direct Teaching Operations, i e., British Council centres where English is taught), and therefore potential raters of M2 in the field. The four scoring procedures used were·

1. *original assessment scale and original method (OM)*
2. *original assessment scale and the profile grid from the second version, i.e., first Profile Method (PM1)*
3. *revised 'global' assessment scale and the global method from the second version (RM)*
4. *new 'profile' assessment scale and the multiple trait procedure (PM2)*

Each team first rated two answers by their assigned scoring procedure as a training session, then they each rated the same ten answers, first giving an individual rating and then agreeing a final rating. The scores of all individual raters and the agreed team scores were collected and comparisons were made within and across teams using Pearson product-moment correlation and the Spearman-Brown prophecy formula.

1.5.2. Results: reliability

Investigation of the scores assigned by the raters as individuals as compared to the scores assigned by raters as teams showed that the original method (OM) resulted in the largest number of rater disagreements (defined as each rater having a different score, i.e., at

least a three-band spread for the three scores): raters disagreed on five out of ten answers; one answer received ratings of 7, 5 and 3. The original assessment scale combined with Profile Method 1 (PM1) resulted in two cases of rater disagreement, the revised assessment scale (RM) resulted in one case of rater disagreement, and the revised Profile Method (PM2) resulted in no cases of rater disagreement. The average single rater reliabilities for the four methods were:

OM	.563
PM1	.864
RM	.883
PM2	.942

Using the Spearman-Brown prophecy formula the reliabilities with three raters are estimated thus:

OM	790
PM1	.950
RM	.960
PM2	.997

(It can be seen that with a more reliable method there is proportionately less additional reliability for more raters.)

On this preliminary check, then, the development of the new rating scale seems to have been of marked benefit to reliability. The use of a multiple trait procedure in the form of both the profile grid (PM1) and the profile scale (PM2) also contributes something to reliability: in the case of the addition of the first profile method to the original assessment, the result is a major increase in reliability, in the case of the addition of the second profile method to the revised assessment scale, the increase in reliability is only slight, and the single rater reliability for RM is, on this sample, more than adequately reliable. It should be noted that false distinctions were drawn for the purpose of the study: raters in the PM1 and PM2 teams scored all answers using the multiple trait procedure without being invited to decide whether or not the answer showed a 'marked' profile; raters in the RM team were not

offered the possibility of using a PM procedure even if they thought an answer showed a 'marked' profile. Because insufficient raters were available, it was not possible to include a team scoring by the second version scoring procedure and the original assessment scale but without using the profile grid, i.e., by the first Global Method. A full study is required using British Council/UCLES data to check comparative reliabilities with a sufficient data set.

In the British Council context, as explained above, the practical reality is that M2 is scored by a single rater, often working in considerable isolation from other raters. What must interest us in this context is a high single-rater reliability rather than any theoretically but not operationally achievable multiple-rater reliability. For this purpose any of the methods except the original one is acceptable. The investigation of the subsidiary research question in Chapter 4 found an average single-rater reliability of .682, which is considerably higher than the .563 found here. It was hypothesised in discussing the reliability level in the investigation of the subsidiary research question that the reliability level achieved was the maximum achievable with the original scoring method because of the high quality raters used. The finding here appears to support that hypothesis.

1.5.3. Results. validity

The correlations between the four methods were generally quite high. Below are the correlations between the aggregate scores for sets of logical 'pairs':

OM/PM1	.908
RM/PM2	.920
OM/RM	.845
PM1/PM2	.929

The other two correlations are .827 for OM with PM2, and .864 for PM1 with RM.

All these correlations are high enough that we can say with some confidence that all the scoring procedures are getting at the same underlying aspects of the writer's proficiency The lower correlation for the original method with the fully-developed multiple trait procedure, .827, conforms to our perception that these two procedures are at opposite ends of a continuum of procedures from free impression marking to carefully specified analytic scoring: they are the least similar pair of procedures used to score M2Q1. The highest correlation is for PM1 with PM2 (.929). These two methods are very similar in allowing the rater to treat each essay as a multiple sample: conceptually they share a view of writing as (at least potentially) multidimensional. We may hypothesize that the profile grid, although it was without any descriptions for the different criteria at each level, achieved what had been intended simply by allowing the rater the 'space' in which to respond. PM2 takes this much further than PM1, but it may be more an administrative convenience than anything else, since the descriptors are already present in the global version of the revised assessment scale: all the profile version of the scale does is break them up conveniently. The rather low correlation between the two global procedures, .845, suggests that something rather different occurs when a rater works from the revised global scale then occurs when she works from the original scale. Since the scale was revised to be a combination of all the traits of the profile scale within one band level description, we can see the revised global scale as closer to the profile grid and the profile scale than to the original scale: the pattern of correlation levels confirms this view.

The high correlation for RM/PM2 is an important confirmation that these two related methods are yielding comparable scores, which is essential when two methods are used as alternate possibilities with the same set of candidates. We are justified within the limitations of generalizability of this small study in treating scores obtained by the RM and PM2 procedures as parallel scores. The correlation for OM/PM1, while not quite as high, is similarly at a reassuring level, and on the basis of

these data it would appear that the correlation is lowered by the poor reliability of OM. We do not, of course, yet have data to show whether similar correlations will be achieved in operational use.

1.5.4. Procedural effects on score levels

It is worth noticing that the choice of scoring procedure appeared to have a slight but noticeable influence on the resulting score level: Table 5.1.1 shows that OM tends to be more generous than the other three methods, and that PM2 tends to be more stringent. It would appear that as the scoring method has been refined and become more rigorous, it has also become more stringent. When averaged, these differences are quite small, but on a single-rater procedure any differences may be very dramatic for any one individual. This makes it all the more important that the trend in the development of the methods has been towards increasing reliability. Figure 5.1.15. shows the aggregate score for each answer for each scoring method:

Figure 5.1.15.: Aggregate Scores: Essay x Scoring Method

	Method			
	OM	PM1	RM	PM2
Essay				
1.	6*	7	7	7
2.	7	7	7	6*
3.	5	6*	5	5
4.	6	5	6	5
5.	8	8	7	7
6.	4*	3	2	2
7.	4	4	5	5
8.	4	3	4	3
9.	6	6	5	5
10.	5	5	4*	5

The single asterisk * indicates where there is an aggregate score which is noticeably different from the others: however, even in these cases the 'wild' score is only different by a single band (e.g., No.1: ~~7*6 6 6~~ 6*7 7 7). The

widest range of aggregate scores on the different scoring methods is three bands (e.g., No.6: 4*:3:2:2).

1.5.5. A SAP procedure?

In discussing the development of the scoring procedures for M2Q1, little reference has been made to the question of SAPness/GAPness This is a reflection of the reality as it occurred.

The original scoring procedure, as we have seen, made no reference to the intended discipline-specific nature of the writing test, provided scoring criteria only in the vaguest of terms, and in no way indicated what qualities of M2Q1 answers might distinguish them from answers on other, discipline non-specific, writing tests. Among the problems with the original version of the scoring procedure reported by the British Council or observed by this researcher, issues of the non-SAP nature of that procedure did not appear. Similarly, none of the raters involved in the piloting of either the first or the second versions remarked on the procedure's suitability or otherwise for its purpose: we shall consider the evidence of whether or not raters treated the second version as a SAP procedure in the next section. The study described in Chapter 4 took place before the development of either the first or second scoring procedures - was in fact the impetus for those developments - and applied the first scoring procedure to all three tests, the two SAP questions and the GAP question alike. None of the raters in that study remarked on the suitability or unsuitability of the original scoring procedure for scoring either SAP question or the GAP question, although they did remark on its generally unsatisfactory nature.

We saw in sub-section 1.1.2. that the only guidance in the original scoring procedure which might speak to the issue of discipline specificity in the question/answer was problematic. First, it was suggested that candidates should not be penalised for "making factual

errors in a subject with which they may not be familiar" (ELTS Administrators' Manual, p. 5). Seaton (1983) suggested that in scoring M2 the language is important and the content is not: this view, and the suggestion that raters should ignore content errors, removes from the writing test the most obviously discipline-specific characteristic It also increases the raters' problems with irrelevance, as discussed in sub-section 1.2.4.2. This researcher's small survey of faculty at the University of Edinburgh, in which faculty were questioned about the criteria applied by subject specialists when evaluating writing in their discipline, showed that content was primary. From the 24 faculty responding, all of whom were highly experienced with overseas postgraduate students, there were 49 descriptions of criteria relating to content (e.g., factual accuracy (15); demonstrate understanding of the subject (5); factual relevance (13); no extraneous material (4)). The next most frequent group of responses were 'rhetorical' (e g., cogency (6); coherence (7); completeness (2); logical development (4)). there were 29 responses in this category.

Clearly, if the question and answer were truly discipline-specific, and content criteria were primary, raters of the kind used for rating M2, i.e., language teachers of varying degrees of training and experience, would be unlikely to be able to judge the relevance of the answer to the question since they would not have the necessary background knowledge. Nevertheless, in denying the operation of a content criterion, the M2 scoring procedure loses a good deal of the validity it could have, and which many testees and score consumers assume that it has. Also, the instruction to ignore content leads to the problems for raters over plagiarism and irrelevance which have already been alluded to. Although a procedure for dealing with plagiarism had been developed, it remains to be seen through detailed monitoring of the scoring procedure that plagiarism is now handled satisfactorily. Piloting of the third version had not shown that problems over the handling of irrelevance had been disposed of.

When we consider the criteria which are applied to the scoring of ELTS M2Q1 we find little to suggest that the scoring procedure is restricted to, or particularly suitable for, a SAP writing test. We can see that 'communicative quality' has no SAP character. Argumentation and organisation are conventional terms from modern rhetoric commonly applied in the teaching of 'expository' writing, i.e., writing for interactional, primarily instructional, purposes, as we saw in Chapter 3, Section 4. While some of the research reported there suggests that there are discipline-specific forms of argumentation and of organisation, there is nothing in the criteria as stated for M2Q1 to suggest differences in the way they are applied to the different Modules. In fact, the 'argumentation' criterion suggests a de-emphasising of expository characteristics and a move toward personal writing:

> *Although 'originality' is restricted by the kinds of questions asked, each candidate should be able to bring in some new information or personal reaction, and to relate to the topic of the question in an individual way. (ELTS: Assessment Guide for M2 Writing, 1985, p. 3)*

Similarly, while research suggests that there are some specific linguistic features of register in different disciplines, nothing in the scoring procedure for M2Q1 indicates that raters should look for discipline-related register variations in candidates' writing.

At all stages of development the M2 scoring procedure appears to be discipline non-specific. The third version, with its clear specification of traits of organisation and argumentation, which at least partially correspond to rhetorical structure features valued by subject specialist faculty, approaches suitability for the scoring of academic writing, but it is not discipline-specific. Indeed, if the procedure were discipline-specific it would be necessary to have a separate procedure for each Module, perhaps for each question, as happens with the M2Q2 procedure (Hamp-Lyons, 1986). Our 'SAP' writing tests, as scored in this study, cannot be considered to be so on procedural variables.

2. RATER VARIABLES

On a writing test, like M2, which is scored by only one rater, differences between raters' behaviours for whatever reasons will result in the increased probability of unreliable scores. We saw in section 3.3 of Chapter 4 that some raters showed greater differences in rating behaviour across questions than others. We also ended the previous section with a discussion of what, if anything, makes the scoring procedure at any of its developmental stages a SAP rather than a GAP procedure, and concluded that there is nothing in the scoring procedure per se, either in the version used for scoring the data in Chapter 4 or in the second and third versions, upon which to base such a distinction. If the procedure is not SAP, it becomes even more critical to understand what it is that raters are actually doing that is general, general academic or specific academic in nature, and to understand as much as possible about the rater variables which are operating and the degree to which these are related to issues of the SAPness or GAPness of rater behaviour. In this section the tape recordings collected during the research and development of the scoring procedure are studied in detail for the understanding they can provide of how individual raters respond and why.

2.1. Introduction to the ethnographic study

2.1.1. Rationale

Research into essay scoring appears to have neglected a consideration of the paradigm within which research in writing now centres itself. that is, that writing is a process of interaction with a reader, just as reading is a process of interaction with a writer. In this paradigm, we can only look at the composed product of a writer through the eyes of a reader. Writing researchers have recently tended to proceed from the position that the writer is the only valid reader of her writing, and that the only valid insight into the writing process is through her eyes.

In such a view, protocol analysis, oral (e g., Hayes and Flower, 1983; Jone-, 1982) or written (e.g., Hamp-Lyons, 1985) becomes the only research method for the investigation of writing as a creative and communicative activity.

However, if one accepts that it is also possible to learn about what, how, and why the writer writes through the study of composed products, then other research methods and informants are available. In such a view, process studies of essay raters offer a rich source of data for understanding how readers respond to writing and how writers respond to their sense that there is an audience 'out there'. Although as we saw in Chapter 2, Section 3, there have been a number of ethnographic studies of writers (e g , Emig, 1971; Perl, 1979, Sommers, 1980, Jacobs, 1982; and in ESL, e.g., Lay, 1982, Zamel, 1983; Heuring, 1984), and a smaller number of studies of essay scoring procedures which pay some attention to what raters do (e.g., Diederich, French & Carlton, 1961; Hake, 1973; Robinson, 1985), we do not yet have a body of ethnographic studies of raters.

2.1.2. Design

The study reported in this Section is based on the taped material collected as part of the development of the second version of the M2 As essment Guide, and springs from the belief that through detailed observation of raters in action it would be possible to better understand what raters actually do when rating, how, and why. These data were collected during an extended rating session, with four raters, which formed part of the piloting of the second version scoring procedure. The session lasted all day, with rests and lunch breaks, and during the whole session the researcher was present and guiding the proceedings, but not guiding raters' judgements. The full piloting procedure is reported in sub-section 1 2.5. of this chapter. It must be noted that none of these raters were the same raters as those who rated the answers used in the study reported in Chapter 4. The answers used were M2Q1 answers, a sub-

set of the original data set of 126 cases. The raters were all qualified and experienced teachers of English as a foreign language, who had all previously marked similar answers (Figure 5.2.1).

FIGURE 5.2.1.: The Raters

Rater	Nationality	Approx. Age	Qualifications	Years of TEFL Experience	Overseas Experience
A	British	35	M.A.TEFL	8	S. America Middle East
B	British	28	M.A.TEFL	3	Japan
C	British	30	M.Sc.AppLings	4	None
D	British	43	M.A.TEFL Ph D.	10+	Middle East India S. America

Figure 5.2 2. summarizes the language backgrounds of the writers, and their language levels in terms of overall ELTS scores. The answers were not seen in the order in which data are presented here. raters saw the papers in randomized order. The writers' names and other identifying information were removed from the scripts, so that the raters had no information about the writers other than what they extracted from the essays.

During the rating session, raters were given a fixed amount of time in which to make their independent decisions about each answer in turn. Each rater then reported his score and briefly described the features of the answer which led him to his decision. The raters as a group then

FIGURE 5 2.2.· The Writers

Writer Score	Language Background	Overall ELTS
1	Chinese	8
2	Chinese	7
3	Chinese	6
4	Chinese	5
5	Chinese	4
6	Arabic	6
7	Arabic	5
8	Arabic	4
9	Arabic	3
10	Spanish	5
11	Spanish	5
12	Spanish	4
13	Spanish	5
14	Spanish	4
15	Japanese	6
16	Japanese	6
17	Japanese	7
18	Japanese	4 5
19	German	8
20	German	5 5
21	Francophone African	7
22	Greek	5.5
23	Indian	9

discu ed their differing perceptions of the answer. All the discussions were audio recorded, resulting in six hours of taped material.

In the study the intention was to·

> *look for me nings from within the situation, allowing the categories for description to be determined by the scene itself. The goal (was) to provide a description that re onates with the members' point of view. Mehan, 1977. 46-47)*

The recordings were studied, both in transcript and directly, for the insights they could offer into the processes and criteria raters were applying when reaching judgements about these answers. Since the main focus of this investigation was not simply a general understanding of

how raters rate essays, but more specifically an understanding of what raters do that makes their rating of 'SAP' essays a 'SAP' procedure, the search was first for SAP-related rater responses. As we saw in Chapter 3,section 4, and as is confirmed by the study of faculty at the University of Edinburgh reported in the preceding section, the principal criteria for evaluating subject-specific writing on academic courses are related to content and rhetorical structure, with content far more important in faculty self-reports. For a SAP writing test, then, we would expect to see content similarly having primary importance in raters' judgements, and the study therefore begins there.

2.2. Rater content knowledge and content effects on scores

The main basis for the claim that M2Q1 (and by intention of parallel design, SAPQ) is a SAP writing test is that the content is discipline-pecific. Since raters will each know varying amounts about the content of different questions, content effects will occur in how much appropriate and accurate knowledge each rater brings to each answer. The degree of importance ascribed to content in answers by the raters is an important indicator of the extent to which the writing test is in fact discipline-specific.

In this sub-section the transcript will be examined for evidence of direct effects of raters' knowledge or lack of knowledge of the subject matter (i.e., the factual information) required by the question or contained in writers' answers on the way they respond to the writing. Analysis of the transcripts shows few references to the content of answers in terms of factual information.

The clearest case of response to content comes in response to the writing of writer 3 (a Technology student later removed from the data set, and not included in the analysis in Chapter 4). Extract 5.2.1 gives part of the discussion of this answer: the full discussion is given in Appendix F:

Extract 5.2.1: Writer No. 3

Rater B: *I wondered too... it crossed my mind... that perhaps it's easier for these people who are lifting... quite... impressive-sounding terms directly from the text, you know... coefficients of various things... so I tended to... then go for a 7 on that... because it is reasonably well-organised and it does directly answer the question. .*

Rater C: *I found it difficult to assess for the reason that I didn't fully understand the content. ...I feel it is a lot easier to write this sort of...sophisticated jargon...*

Rater A. *that's a bit unfair, isn't it? I mean what do they have to do then... if they're writing Technology... to be able to do an 8?*

Rater D: *I do find it not so easy to judge things like... the theme, logical presentation... communicative effectiveness (pause) ... particularly as one... in the other ones tends to think in terms of the accuracy and relevance of arguments whereas we can't tell... I suspect, for example, in line 4 "tough" is not the right word... I would've thought "brittle" is the right word there...*

Rater B: *I think there is a technical usage... I mean I find it relatively easy to understand... because... I did do science to A level and that makes it easier...*

Rater C: *it sounds like a textbook to me...*

Rater B: *but it does... I mean there are technical usages and I think tough is...maybe...I'm not quite sure...*

Rater D: *that makes it difficult for us to assess it...on the basis of what I can... within those limitations... I gave it a 7... I agree that there are a number of... particularly er... I mean, can you say "low hardness" and "high hardness" - is that technical?*

Rater A: *no*

Rater D: *well, how do you know?*

Rater A: *because I've taught ESP and... this area and I'm sure that's not right...*

Rater B: *I think it should be "low hardness value"...*

To varying degrees all the raters are in difficulty here, and they are prepared to accept the fact of that difficulty to varying degrees. Rater C is perhaps most honest in seeing the difficulty of making an accurate assessment without full mastery of the content: however, he also seems inclined to discount the importance of the specialist aspects of the

answer, arguing that "sophisticated jargon" of this sort is "easier to write". He remains fixed on the attempt to judge the language apart from the content, and comments unfavourably on the cohesion. C's final band score is the lowest, at 6. Rater D also admits to difficulty in making a judgement, but relates this to other factors in his process of forming a judgement, emphasising the fact that he can't judge the relevance and accuracy of the arguments in this answer. He comments humourously ".. the fact that I understood it, I was so relieved that I gave him a 7", but has in fact noted both some strong elements and some weaker ones which feed into his judgement. Rater B claims some expertise ("science A level") and claims to find it "relatively easy to understand", but cannot give an authoritative answer as to the correct technical meaning of 'tough'. Rater B justifies his judgement of 7 as final band score by reference to organisation and relevance, and does not bring out his awareness of some content inaccuracies until rater D comments on 'tough' and 'brittle': it would appear that his judgement is little related to the level of content accuracy but rather to other features of the answer and to his assessment of task difficulty. Rater A is most positively affected by this answer and defends it quite strongly, commenting that he found it "very clear"; in contrast to rater C, he does not notice the grammatical inaccuracies until these are pointed out by other raters, but he does in response lower his final band score to a 7. Rater A confidently asserts that 'low hardness' and 'high hardness' are incorrect technical usage because "I've taught ESP ... and I'm sure that's not right". In this situation, having done A level science or taught ESP are claimed as expertise: clearly they are not the kinds of expertise which a subject specialist would claim.

We are left, after examining this discussion closely, with the sense that raters are in different ways attempting to discount or reduce the importance of specialist knowledge, both in the answer and in the repertoire of skills of a rater.

In the next extract (Extract 5.2.2), raters discuss the content of writer number 6's answer, but in a way which brings them closer to structural considerations, particularly to argumentation, than to factual accuracy.

Extract 5.2.2: Writer No. 6

Rater B: *...it doesn't quite run smoothly from the question of the character of the doctor into the...the pros and cons...of...you know...for the advantages and disadvantages ...I thought that would need to be better...generally very easy to understand...*

Rater C: *...didn't think he was a 'Very Good Writer'...there are occasional lapses in logic...I didn't understand the doctors...I mean...unless he thinks you meet more potty porters than doctors, I don't know...*

Rater D: *no - doctors make... potty people...*

Rater C: *that's what I mean...is that what happens?*

Rater D: *...make the jump...a logical jump...*

We see the raters trying to reach an accommodation with the writer's meaning here, to make it make sense. To do this, however, their concern is clearly less with the factual information presented and more with using their interpretive skills to get at the structure of the argument. There is, however, a sense that they are all fairly positive about the answer: this is one of only three answers on which all raters were agreed, scoring it at final band 7.

The discussion of writer number 8's essay also touches on the truth value of the content (Extract 5.2.3).

Extract 5.2.3: Writer No. 8

Rater D: *...theme is straightforward enough... two paragraphs, one illustrating why you need to have medical knowledge being a doctor, and the other paragraph... the value of being a nurse... or a paramedic... but there's not much content...*

...'

Rater C: *... I thought the first two paragraphs were irrelevant and unnecessary... in fact he's got the middle paragraph wrong... he means cannot... and that makes (mumbles of agreement)... and that may be a slip, or*

maybe he's confused in what he's trying to say... I don't think it's clear...

Rater A: *...well... he says... "It is essential to have a good medical knowledge"...*

Rater B: *... I think he's just missed out 'not'...*

' *Ellipsis at the margin indicates part of the transcript has been omitted.*

The omission of 'not' is seem as a serious threat to the acceptability of the answer, and the scores (raters A, B and C give final bands of 4 while rater D gives 5) seem a little low for an answer of which rater B says "...the message does come across... that they've understood the question... and are attempting to answer it". It would seem from this that content must at least be internally consistent, even if factual accuracy in an absolute sense is not a criterion. From an analysis of these tape recordings it does appear that raters were conforming to the advice in the original version of the scoring procedure, retained in the second version scoring procedure, that factual accuracy should be more or less discounted as a scoring criterion.

We find passing references to factual content in the discussions of writers 10, 14, 15, 17, 21 and 23, but these discussions focus more on structural characteristics, especially on the quantity and quality of support a writer brings to the argument being presented, and are therefore considered in section 2.3.

2.3. Raters' responses to argumentation

In the second version scoring procedure, using the Assessment Guide (1985; 1986) 'argumentation' is described as relating to a "meaningful and interesting answer" (p. 3). Examination of the transcript indicated that raters had two main concerns which fell into this category: the kinds of supporting ideas in terms of the details, examples and experiences the writer used and how they were presented; and the relevance of those

supporting ideas to the question asked. We shall look at these concerns separately in the next two sub-sections.

2.3.1. Argumentation

In the first extract (Extract 5.2.4) from the discussion of writer number 21's answer, we see a large measure of agreement among the raters as to how argumentation is functioning in the answer, making a central contribution to the overall structure of the answer:

Extract 5.2.4: Writer No. 21

Rater A: *7 6-5... communicative quality... seem to be some strangenesses in the argumentation that er... things such as... the idea that you're feeding a large population in a small area... and it would take less time to achieve... I didn't really understand why... um... so 6 or 5... 'Competent' or 'Modest'... and I think I'd probably give him a 6...*

Rater B: *I think it's a 6 too... largely because... although it appears that he's got it all neatly worked out and planned with the... again, with the key words in the right places... it's not, I think, worked out systematically enough to warrant some sort of (...) I mean there's... there's no details to that effect sort of laid out... but still... clearly very... nicely written strings in other parts... and... an obvious awareness of what structure should be...*

Rater C: *I agree... 6... hasn't quite mastered all the techniques of structure... structuring for an essay and yet he's obviously halfway there... definitely a competent writer...*

The comment by rater A about "strangenesses in the argumentation" comes close to being a content comment - indeed content and argumentation cannot be entirely separated - but as the discussion unfolds it becomes evident that the raters are prepared to accept the factual basis for the argument the writer wishes to make, if it had only been "worked out more systematically". All three raters (rater D was not present for the scoring of this essay) awarded a final band of 6.

In the next extract (Extract 5.2.5) from the discussion of the answer of writer number 15, we see the raters seeking to negotiate whether the answer is sufficiently well-argued to permit them to admit the relevance of the supporting material the writer offers: that is, the material is or is not relevant depending on the strength of the argumentation in the answer.

Extract 5.2.5: Writer No. 15

Rater C: *I thought it was reasonably well-presented... it was coherent... I didn't think there was an argument running all the way through...*

...

Rater B: *I thought it was quite well structured, and I saw, in terms of criterion argument... that he was using his own ideas there... giving that little story... to support his point of view...*

...

Rater A: *I didn't understand really... what his message was... I thought the organisation, argument was not good... the second paragraph seemed to me to be totally irrelevant (...) the logical structure of it was not very good for me...*

Rater D: *I think... he displays quite a reasonable... well quite good... grasp of the language in... deploying his argument ... bearing in mind it's a divergent type of essay ... even if his example isn't really spot on I think it's an attempt to integrate his experience with the... theme of the thing...*

...

Rater C: *I'd have to be convinced that the language was more important than the structure of the whole essay ... which I think is lacking..."*

Rater A: *Yeah I mean... I'd have to be convinced about the relevance of this second paragraph... it seems to me the argumentation is weak..."*

Rater B: *... I think there's probably a sentence missing, to make that and supporting the decision he's made ... because what he's trying to say with that is that ... er ... people in psychological experiments have thrown up some very interesting effects on people who are in prisons as inmates ... and for this reason it's very interesting to get the point of view of a prisoner - more interesting than somebody who's in... etc... so I*

mean... I suppose what I've done is... filled in a logical gap there... but I think there <u>is</u> a clear structure which perhaps needed something in between to support it. But I feel it <u>is</u> relevant...

...

Rater C: *... I just found the... the supporting detail, if you like, wildly irrelevant...*
Rater A: *Yeah .. that's right... not well structured. .*
Rater D: *I don't really see that there's that much of a gap...*

Rater C finds the answer poorly argued although coherent, rater A not only finds it poorly argued but is unable to follow the writer through the thread of the discourse. The other two raters, B and D, do not have a problem with the answer, recognising that the argument does not follow a conventional academic English rhetorical pattern (rater D actually refers to it as "divergent") but being tolerant of that fact. There is in A and C's comments a sense of 'wrongness' about the discourse: what comes through from B and D is a sense of 'difference'. These differences in how raters are prepared to value the argumentation used are reflected in the scores they assign: rater D scores this essay as band 7, as does rater B. Rater C scores it band 6 and rater A band 5.

In Extract 5.2.6, taken from the discussion of writer number 10's essay, we see the raters having problems arriving at a judgement of the argumentation in the answer because of a mismatch between the quality and the quantity of argumentation in the essay:

Extract 5.2.6: Writer No. 10

Rater A: *... I started off... it had a certain spurious attraction... I started off with a 6-5-4 and then actually looking through it, it seemed that there was basically... you know, one reason repeated about three times in different ways for why... she'd like to read this book (he illustrates)... it's just the same thing being repeated time and time again... and for me... logic and argumentation dropped it...*

...

Rater D: *...when I got to 'first' I was looking for 'second'...*
Rater A: *Yeh... me too...*
Rater D: *..."for two main reasons, first"...der...der...der... I wondered whether 'moreover' was supposed to be 'second'...*
Rater A: *no. it wasn't...*
Rater D: *...but since it's the same point... or virtually the same point...*
Rater A: *yeah, and "the main interesting point" is again the first point...*
...
Rater C: *...I know there's repetition but I mean it is... quite well disguised and it's... she's expressed it differently three times... this is every 'O' level candidate's dream...*
Rater D: *... there might only <u>be</u> one argument... it depends how it's presented... but this is presenting one argument as if was three...*
Rater A: *... it's suggesting that there are more and then .. promising what it doesn't deliver...*

Rater B takes no part in this discussion, but in his initial input (refer to Appendix F) gave no sign that he saw, or was troubled by, the repetitive nature of the argumentation. The writer appears to have been prompted by a sense of what was expected into a reduced version of the classic five paragraph expository theme - introduction, three arguments in three paragraphs, conclusion - but does not have the material to follow this through. The raters' expectations are to a greater or lesser extent aroused by the recognition of a familiar discourse structure; raters A and D are clearly very troubled by the failure to "deliver"; rater C takes a more cynical approach, on the basis of the evidence available to us rater B is prepared to see the answer as "not very well structured" and leave it at that. Rater B does not alter his initial judgement of 6; raters A and C give final bands of 5, while rater D scores it 4.

The five-paragraph theme is hardly a discipline-specific discourse genre: it is typically found in textbooks which teach expository writing for multiple purposes. The genre is intensively taught in 'Freshman

Composition' courses at American colleges and in 'O' level courses in British high schools (hence rater C's reference). It is, then, an essentially GAP genre which evokes this reaction from the raters. This observation of an expectation in the raters that answers will fall into a conventional rhetorical structure of academic discourse recurs in the discussion of the essay of writer number 7 (Extract 5.2.7):

Extract 5.2.7: Writer No. 7

Rater B: *...message is there but er... there's a... it lacks consistency.:. and it lacks... a clear argument... in fact because it's sort of.. slightly circular in that it starts off saying one thing and then says the opposite by the end of the thing... and... not in a way which is acceptable... not sort of 'on the one hand', 'on the other'...*

...

Rater D: *.. relatively easy to understand... with occasional odd... sections... he marshals a convincing number of points in er... in favour... some overlap... but er... convincing...*

Rater B has misinterpreted a point of content in the answer, and rater C clears this up for him. Rater B's comments suggest that he is prepared to accept a circular argument if it is conventionally marked, which in turn suggests that he is valuing the conventions through which the message is conveyed over the real value of the message itself. It also shows how closely tied together argument and organisation are in the rater's judgement.

2.3.2. Relevance

We saw in sub-section 1.2.4.2 that there are two kinds of reasons why an answer may be irrelevant: the treatment of the 'stock answer' was changed in the third version, as we saw in sub-section 1.3.6., and there was in this sub-set of the data only one answer which some raters considered to be a 'stock answer', one which was of so low a level it made little

difference in terms of score assigned. Throughout these discussions, raters make no reference to failures to understand the question or the input text, although such failures may in fact lie behind some of the answers and raters did not recognise this. The second kind of reason includes the 'challenge to the question', which is dealt with in Chapter 6 and is not discussed here, although references to challenges appear in the extracts which follow; it also includes cases when the candidate is unable to answer the question.

We are left, then, with those discussions which simply acknowledge the irrelevance of an answer or part of an answer without attempting to addre s que tions of why it is irrelevant; and with those discussions where raters do not agree about the admissibility of an answer. Extract 5.2.8, from writer number 5, is a good example of this:

Extract 5.2.8: Writer No. 5

Rater D: *...there is a problem of course in that he doesn't answer the question ..*
Rater B: *I wondered when you'd get to that...*
Rater A: *Thank you...precisely...*

Rater B: *As a scientist how would <u>you</u> defend...*
Rater A: *How can you give him any mark at all? He hasn't answered the question...*

Rater A: *well, I said it was irrelevant... he didn't answer the question at all... well I mean... whatever we give for a totally irrelevant answer... which is what - 2 or 3? ("2") well, this is totally irrelevant... doesn't answer the question at any stage...*

...

Rater D: *well it's not entirely irrelevant in that it does suggest or begin to suggest methods for reducing...*
Rater B: *but that's still not the point of the question...*
Rater A: *yes but he doesn't defend it... he's supposed to be defending... thecontinued use of such potentially harmful processes... on what basis, you know... that you're benefiting mankind or you're...*
Rater C: *I gave it a 4 anyway... I didn't realise he hadn't answered the question...*

We see that rater D was not disturbed by the failure of the writer to select content for the answer which relates to the exact topic of the question, and is prepared to consider the answer admissible, raters A and B, however, react very strongly against it. Rater C was presumably not disturbed by the irrelevance since he had not noticed it; we do not know why he rated the answer so much lower (4) since he takes no part in the discu sion, which centres wholly around the issue of relevance. These differences in the three raters' perceptions of the admissibility of the answer are reflected in the scores assigned: rater D gives a final band 6, raters A and B give final bands of 4.

In the next extract (Extract 5.2 9) we observe an awareness on the part of all raters that writer number 16, an architect, is trying to integrate his own knowledge and expertise with the topic set him in the question, but they differ in their willingness to accept his attempts as admissible:

Extract 5.2.9: Writer No. 16

Rater A: *He's saying... "I don't really know very much about this and I will look at it from my... point of view which is that of an architect".. and he's desperately trying to find something in there that he can hang his architectural experience on to .. and he's managed that quite ingeniously I think...*

...

Rater B: *I think his profession is... in the way it's been worked in... is a little bit irrelevant... well, it's irrelevant I think... because it is not, for me, convincingly argued that... the things he wants to know... are actually of interest to an architect... the conditions, what prisons are there, and what kinds of rooms there are.. is just not a convincing argument at all*

...

Rater C: *...it doesn't answer the question... no way could it answer the question because... he doesn't say... he chooses talking about architecture but he doesn't say*

why that is related to reading one book rather than the other...

...

Rater D: I don't think we can say he hasn't answered the question (...) I gave it a 5... the language is not bad and... the question only says 'Choose one and give reasons for your choice'... and he's given what seems to me to be a very lucid although possibly ridiculous to us reason...

Raters A and D accept that the writer has made some attempt, however feeble, to answer the question from within his own disciplinary knowledge, and give him credit for making the attempt even though it is not very successful; raters B and C are not prepared to do so This is reflected in scores of 5 from raters A and D, 4 from rater B and 3 from rater C.

Other discussions of the relevance or irrelevance of answers occur, notably concerning the essays of writers 8 and 15, both of which have been extracted in earlier sub-sections. From all the occurrences, it would seem that rater C is generally rather strict with irrelevance once he recognises it, while rater D is quite lenient; raters B and A seem to be more varied in their response from case to case.

2 4. Raters' responses to organisation

The Assessment Guide (1985;1986) characterises organisation as the "logical structure" of the answer and describes it in terms of cohesion and coherence features and of paragraphing conventions. Many of the discussions link organisation and argumentation closely, as we saw with extracts 6 and 7: this linking is also shown in the results of the faculty survey described in Section 1, where faculty frequently coupled coherence and cogency, or organisation and logical development.

In Extract 5.2.10 we see that features of organisation are the primary characteristics that raters are responding to in the writing of writer number 11:

Extract 5.2.10: Writer No. 11

Rater B: *WhatI'm worried about is that there's some suspect logic... in it... it's veiled... there's a certain... neatness of things which makes you think there's a logical structure 'cos they've in fact put in various key words but... I don't think it's quite there (... reads from Assessment Scale) "inadequate connectors and cohesive features"... in fact those are precisely the bits that... are good...*

...

Rater A: *I - I got a message, but there was this sort of vacillating message going through it... and I felt it wasn't as organised .. as well organised as it could've been... and there were... sort of... linguistic .. er inaccuracies and appropriacies...*

Rater D. *...it's definitely a 5 as far as the message is concerned (pause) and it doesn't quite fall to a 4 (. .) but how far do you mark it down for linguistic inaccuracies?*

Rater A: *well according to the system you can't mark it down very much for those inaccuracies. . you choose the lower of the two bands you've gone with...*

Rater B: *I'm convinced it's because they've got these... cohesive devices a lot...*

Rater A: *...they've been taught them...*

Rater B: *...they've been taught them... you know, the 'in conclusion' and the 'of course' and the 'on the other hand'... it sounds so well structured... and yet I mean in fact... I think it's slightly suspect...*

We cannot help noticing that the raters are now finding fault with the conventionalised rhetorical techniques the writer is using, when they earlier (in discussing the essay by writers number 10 and 7) criticized the writers for not displaying these conventional features It is difficult, finally, to get a sense of what the raters want from a writer in this regard. However, while the raters seem to be agreed that the essay is somewhat difficult to score because it does not demonstrate an even performance (in fact, what came later to be known as a 'marked'

profile), they are also agreed about the score, all scoring it a final band 5.

In Extract 5.2.11 the issue is a little different. discussion centres on changes in the level of performance of writer number 17 through the essay rather than from criterion to criterion:

Extract 5.2.11. Writer No. 17

Rater B: *I started 4-5-6... 6 is out... it's not... it takes a couple of readings, at least, for me... to get the message properly (...) generally there is an organisation there, and an argumentation... um... even though it's slightly hard to work out... but I think it fits quite well... into what I would call a 5...*

Rater C: *I think it doesn't read very well... on my first reading I gave it a 4... I then read it again a bit more carefully and decided that I would up it to a 5... because of vocabulary and... possibility of structures, even if they're not quite there... and the meaning comes across...*

...

Rater D: *... I can't give it more than a 4... it seems to me to be... he sort of gets to the end of his... first paragraph or whatever... second paragraph, isn't it?.. actually third paragraph, isn't it - the way it's written... and he seems to have run out of steam, and I think that's it, I don't think you'd ever get any more 5 level... steam out of him...*

Rater A: *...I still don't know what the message is... I'm afraid... I... I mean... he starts off "the following reasons" and... there <u>are</u> no reasons...*

While the raters never explicitly centre their discussion on organisation, the sense in reading this extract is that a weak organisation is intruding into their ability to negotiate the other difficulies they have with the text. Rater C found it easier on subsequent readings, which in itself suggests an opaque text structure; rater D's wrestling with the number of paragraphs indicates that the paragraphing conventions do not conform to his expectations based on his understanding of the conceptual divisions of the answer. In their scores, raters seem to be divided as

to whether a closer reading improves their perception or merely reveals more problems: raters A and D award final bands of 4 while raters B and C award final bands of 5.

There are other discussions of organisation, but in general these do not occur in isolation from discussion of other criteria: examination of the transcript does not suggest any features of organisation which we could relate specifically either to discipline-specific characteristics of organisation in writers' treatments of their answers to these questions, or to raters' expectations of such characteristics in writers' responses. We find only the references to typical cohesion markers found in expository writing. We do not, then, find any evidence in the raters' application of an 'organisation' criterion to suggest that they are responding to this as a SAP writing test.

2.5. Effects of linguistic features on raters' judgements

References to linguistic features are seeded throughout the discussions. In some cases discussion of linguistic features seems to be central, but more often linguistic features are mentioned as a stage in the judgement process. We observe a variation in whether linguistic features are mentioned late in the explanation, to support a judgement about the selection of one band rather than the other from a two-band range, or whether they are mentioned early, presumably because linguistic aspects of the answer played a fairly important role in leading the rater to a certain band range. Use of linguistic characteristics in judgements in the latter way conforms to the use faculty suggested they made of linguistic criteria in some of the comments on the faculty survey referred to in Section 1, and to the findings of Bridgeman and Carlson (1983). The former use appears to run counter to reported research into the criteria for judgements applied by subject specialists.

Extracts 5.2.12(a) and 5.2.12 (b) are both from the discussion of writer number 3, but show different approaches to a judgement:

Extract 5.2.12(a): Writer No. 3

Rater B: *I started with 7-8-9... there are certain... um... areas where... he hasn't used anaphora in the right way, for example, or he - or she - ... hasn't included... the two parts of the sentence there should've been an 'also' to make it balance properly... maybe the slip-up on 'reason' plural - this kind of thing... just a... number... of little things, though otherwise... it reads very well... I wondered too... it crossed my mind... that perhaps it's easier... for these people who are lifting... um... quite impressive-sounding terms directly from the text, you know,... coefficients of various things... um... so I tended to... then go for a 7 on that... 'cos it is reasonably well-organised and it does directly answer the question...*

We can reconstruct rater B's judgement process as a decision to narrow the range to 7-8 on the basis of some flaws at the linguistic level, followed by a further decision to select a final band 7 because of reservations about difficulty level which were discussed in section 2.2. Rater A's approach to a decision on the same essay is different (Extract 5.2.12(b)):

Extract 5.2.12(b): Writer No. 3

Rater A: *I think it is very clear.. the message is very clear... um... it's well argued, there are some nice... er... anaphoric, are they? references... "hence, aluminium"; "which has a low density"... and I ended up giving it an 8...*

Rater A does not tell us what three-band range he started with; however, he clearly appoached the decision from the communicative end of the spectrum of criteria, rating the answer very high on that basis, and then found support for his decision as he moved to argumentation as a criterion and finally to linguistic characteristics.

We see in the discussion of writer number 13's essay (Extract 5.2.13) the use of linguistic criteria to guide a choice of bands where the only other apparent criterion raters are applying is length of sample:

Extract 5.2.13: Writer No. 13

Rater B: *First of all there's really an inadequate sample,,, the second thing is that the first three lines which form most of the thing seem to me to be a restatement of the question... and therefore the sample is reduced even more... and the bit that then seems to be his own production is extremely poor... says virtually nothing and um... with many inaccuracies*

Rater C: *I thought about 4 for a while... just because of the length I think... but on closer inspection it was not worth it... a 3*

Rater A: *... it just... seems it's got better control of the language than a 3 would suggest...um... the message I can work out very easily... one or two mistakes but certainly fewer than you would expect.. for a... you'd expect more mistakes for a band 3..*

Rater B: *..."whether we know it, that processes are more or less only dangerous for us... as if they are also helpful and necessary"..*

Rater A: *"or if*

Rater B: *"or if"... yes... "very urgently"*

Rater C: *I honestly still don't know what he wants to say...*

Rater D: *I think the only real error there... OK... 'h' missing out of 'whether', not significant... "that processes" as opposed to "those processes"... the rest of it is unusual but makes sense... "are more or less dangerous for us or if they are also helpful" - if you use a different intonation... it makes... not only, but also... structure...*

Rater A: *...he's... yeah...*

In this case it is primarily on linguistic grounds that rater D assigns band 5 and rater A band 4; raters B and C have the same reason but different perceptions, and both assign band 3. Extract 5.2..14 shows a different use of linguistic criteria in the discussion of writer number 19's essay:

Extract 5.2.14: Writer No. 19

Rater A: *I should... because there are no or... I can't see any... linguistic inaccuracies, come up with a 9... I'm very loathe to do this... so in fact I've given it an 8... now I'm not sure that my reasons for not giving it a 9 are very good ones but anyway that's my (...) the message is very clear .. um... the argumentation is very clear... um... and... there are no linguistic inaccuracies... I suppose by all accounts it should be a 9, I'm giving it an 8.*

Rater D: *... I probably didn't pick up on one or two basic errors in it first time... I don't like the beginning... I think the beginning is a very weak beginning "First I wish to state"... I mean - why didn't he just state it?... Secondly there's a repetition, a serious repetition... in that... (inaudible) ...we start off "The outcome of my choice depends on what kind of material" and then just a bit further down "My choice between them depends on (inaudible) to study" ...this seems to... you know... it really screws up that paragraph I think... it's true that the... the vocabulary and the... the language the linguistic accuracy and so on is very good and so on but er... basically from that point of view I think it's seriously flawed in that first paragraph... so I'm not clear on themessage.. as clear as I could be...*

Rater B: *... my reason for dropping from 9 to an 8 is that really I would expect the conditional.*

The situation here is the reverse of that in Extract 5.2 13, where linguistic features were used to justify raising the score. the raters appear to be applying very stringent criteria to the answer in justifying their decisions not to assign a band 9, and in so doing appear to be implicitly and at times explicitly disregarding the essay's linguistic strengths.

We saw in the extract from the discussion of the essay of writer number 3 in the content section (2.2) that the use of specialist vocabulary was dismissed as 'jargon' by most of the raters; the questions on M2Q1 offer relatively few opportunities for specialist vocabulary, as we shall see in Chapter 6. Specialist vocabulary would be the most obvious way in which the answers would show discipline-specific linguistic characteristics;

there does not seem to be anything in the discussion of linguistic characteristics in the transcript to distinguish the criteria the raters are applying to the scoring of these essays from the criteria they could apply if they were rating GAP essays.

2.6. Evidence of influence of other variables

2.6.1. 'Message'

The expression 'message' recurs throughout the discussions, for example, in No. 8 "the basic message comes through"; in No. 10 "it's very clear, I think, to read... the message comes across"; in No. 11 "it's definitely a 5 as far as the message is concerned..."; in No. 12 "I get some message from her but..."; in No. 15 "I didn't understand really .. what his message was..."; in No. 17 "... it takes a couple of readings, at least, for me... to get the message properly..."; in No. 20 "I think there's a very intermittent message here..."; in No. 23 "...it seemed to me that I was getting the message (...) it was a bit difficult actually getting it through" and from another rater on No. 23 "I can see the problems with it but the message is clear...". Occurrences are many and often appear very early, in the first rater's explanation of the reason for his score. Raters begin with it and come back to it: it is possible for one or two raters to have 'got the message' and for others to have missed it. The extract (Extract 5.2.15) from writer number 17, which begins well into the discussion, illustrates this:

Extract 5.2.8: Writer No. 17

Rater A: *I... I still don't know what the message is... I'm afraid..I mean... he starts off "the following reasons" and there are no reasons... and this last paragraph I just don't understand...*

... (Raters C and B attempt to explain)

Rater A: *I got the feeling that it was a challenge, but I couldn't understand... I still don't understand... this*

bit below the crossing-out... I just... I don't know what is happening there...
... (further explanation and discussion)

Rater A: ...but I mean the mere fact that we're actually having to discuss what the message was... (yeah) ...would suggest to me that the message is not coming across too well...

(at the end of the discussion)
Rater A: I still find it very puzzling"

The term 'message' is used in the original M2 assessment scale which raters were working with, at band levels 5, 4, 3, and 2. It also appears in the description of 'communicative quality' as a criterion in the second version Assessment Guide which these raters were piloting. It seems to have struck a chord of some kind with them, to be a word which expresses a range of responses which are important to them but perhaps difficult to characterise more precisely.

Although it has been included in the 'content' sub-section of this section, 'message' is clearly composed not only of content but also of structural and linguistic features. We often see it juxtaposed with other criteria, for example, in No. 4 "I think the message comes over, but it is so full of inaccuracies and... um... mistakes...", in No. 6 "...message comes through clearly... nicely thought out and argued... it's got good structure... there are few, if any, inaccuracies or inappropriacies..."; in No. 8 "the basic message comes through, but there's not a lot of er... argumentation... and I think then... the linguistic accuracy and appropriacy pull it down even further..."; in No. 16 "I don't have difficulty working out the message... it's fairly clear... there isn't inaccurate vocabulary, really..."; in No. 19 "...the message is very clear... the argumentation is very clear... there are no linguistic inaccuracies ..". Taking all the occurrences and looking for the common thread, it would seem that 'message' is used synonymously with 'communicative quality' as this is defined in the second version Assessment Guide.

Some insight into the meaning of 'message' to the raters may be gained from a closer look at an extract from the writing of writer number 14 (Extract 5.2.16):

Extract 5.2.16: Writer No. 14

Rater D: *I think I feel adamant on this... I don't see on what you can base... more than a 2...*

Rater A: *well I'm sorry... I think there's an argument for giving it a 5... as Rater B recognised...*

Rater D: *"theme" - there is no theme... "logical" - what logic?... there isn't anything there...*

Rater C: *well I don't think it's true that, in 2, for example, that "there is no real communication with the reader having constant problems in making out any message"...*

Rater A: *no... you don't have constant problems to make out the message..*

Rater D: *well there is no message... the message is only started...*

Rater A: *well OK, but there's still a message... you can't say a message is started...*

(This discussion continued at length and concluded with an agreement that the answer should be treated as an inadequate sample.)

It would seem from this, and from all the other occurrences, that 'message' relates to the expectations each writer sets up in the reader in the opening stages of the answer, and the extent to which she then fulfils them. These expectations are different from writer to writer, perhaps depending on some sense the rater gets of the promise of the opening one or two sentences, and they are clearly different from rater to rater for the same answer.

2 6.2. Rhetorical structure and raters' responses to rhetorical transfer

We saw in Chapter 3, Section 4 that English academic prose discourse is expected to conform to highly conventionalised rhetorical norms, and that the expectations of the expository essay, especially the test essay, as a genre are particularly strong. We may predict that such expectations will be brought to a SAP writing test, and specifically to M2Q1, by writers and readers who are fully initiated members of the discourse community. We can expect raters of a SAP writing test to look for and to value responses with a genre-specific rhetorical structure.

Further since, as we saw in Chapter 2, section 4.3, rhetorical structure varies across languages and cultures, we may expect that some writers will display other kinds of rhetorical structures. Different raters may respond differently to the rhetorical structures displayed by writers according to their experience as readers of the writing of other rhetorical communities in general, and according to their personal experience or lack of experience of the language and culture which con titute the discourse community of each writer.

There are several instances in the transcript which suggest that raters are responding to cross-cultural transfer The most striking is in the case of writer number 15, a Japanese at an overall band 5 (Modest User), part of which discussion appeared as Extract 5.2.5 above. The extract indicates that two raters find the answer unsatisfactory, while rater B finds it quite acceptable and is able to interpret it at some length for the other raters. We know from research that Japanese, in oral narrative (Clancy, 1980) and written discourse (Kojima and Kojima, 1978; Hinds, 1983) are able to use referential choice and sentence-position of referents to make their texts cohere in ways not available in English: this is what makes haiku so difficult to translate. We can postulate that such differences of coherence properties predispose, or at least permit, a different rhetoric. Rater B has taught in Japan and speaks and

writes above intermediate Japanese. He actually said, later in the discussion: "Is it a Japanese? Yes, I could've told you. I'm used to ... knowing the Japanese ... certain cohesive gaps". He is aware that the essay breaks the rules of academic discourse in English, but is able to read it as if it did not. Rater D, the most experienced of the raters, is also sympathetic to this piece of writing.

Another instance occurs in the discussion of the answer by writer number 23, an Indian at an overall band 9 (Expert User), as shown in Extract 5.2.17.

Extract 5.2.17: Writer Number 23

Rater B: *"I found it ... just going on and on and not coming to a ni ely rounded proposition or whatever... so I found it difficult to understand and I think it's on the communicative level..."*

...

Rater C: *"... the vocabulary is very impressive at first... you think he's saying something and... I don't think he is!"*

Rater D: *"The structure is pompous but it's clear ... it gives you the advantages first and then the disadvantages. The vocabulary is a bit ... over-expressive, but I don't think you can penalise that (...) it's unfair to penalise him on the type of vocabulary he uses (...) the message is clear, if tendentious .."*

...

Rater A: *..."the argumentation, the organisation, was a bit obscure at times, it was a bit difficult actually getting it through ... He clearly has a nice grasp of the language..."*

...

Rater D: *"Most raters would probably be seduced like I have been ... by the bombast.."*

Rater C: *Well - that just put me off entirely. I thought, anybody who can write that sort of thing..."*

Rater D: *"But that's just a cultural thing..."*

The discussion seems to focus on two areas: the way the language is used at the lexical and syntactic levels, with long, complex sentences and what is often called 'flowery' vocabulary; and the structure of the argument. As the extract shows, rater B finds it difficult to understand; rater C agrees; rater A approves of the language but has difficulty with the argument; rater C finds it perfectly clear, and in fact reconstructs it for the other raters, but has ambiguous feelings about the language. Raters A and D (who both have a background of Classics) respond most favourably to the language the writer uses, rater A considering that the writer has a "nice grasp" of the language, and rater D being "seduced" by it. However, rater A, who has not had significant exposure to students using Indian English (by which is meant the languages of the sub-continent rather than any specific language) has problems with the argument, whereas rater D, who has taught in the sub-continent, does not. Although no mention is made of the writer's linguistic or cultural background in the discussion, there is a sense that some at least of the raters have it pinpointed. rater D says "...I think in the context in which they're working . where doctors are often very much detached from what actually goes on there...prehaps he's emphasising the "able to explain", that is, that you've got the touch of the common people... but it's obscure, isn't it?"

Kachru (1983) discusses Hindi rhetorical structure and gives examples which seem to share the highly embellished character of the advanced level writing of the writer in extract 5.2.17. Kachru does not comment on this apparent embellishment but she does discuss some major syntactic differences between English and Hindi, pointing out that Hindi may appear to have a more nominal style than English, and she says:

> *Indian English texts present difficulties to a native speaker of English because of deviant coherence (e.g., paragraph structure) as well as cohesive strategies (e.g., use of tenses, linkers, lexical sets). The Hindi texts present the same kinds of difficulties to a native speaker attempting to learn Hindi. An increased awareness of different strategies of coherence and cohesion would*

certainly facilitate interpretation of texts in a second language or variety. (65)

The full discussion of this answer makes clear the distaste that raters B and C have for the level of diction, the 'flowery' language. Although rater D accepts that his initial favourable reaction is inappropriate when challenged by B and C, he appears to continue to feel that the rhetorical structure is not only there, it is there in essentially the same form as it would be in standard (academic) English. He is able to see that the intrusive diction is just "a cultural thing". Rater D scored the essay as a band 7, rater B scored it band 6 and rater A band 5; rater C scored it band 4.

And finally, an example from the answer of a Chinese writer, writer number 4, whose overall ELTS score was band 4 (Limited User).

Extract 5.2.18.: Writer No. 4

Rater D: *...it's long enough for a 5 (laughing)... but short points... if you ignore the numbers you see. . there's no connection between the sentences...*

Rater A: *no... but that numbering is a way of doing it... I mean we don't disqualify people for doing it... very nice...*

Rater C: *... it's a study skill, isn't it... if it impresses people so that they don't actually look at the internal logic there...*

Rater A: *yeah... you can't knock that...*

Rater D: *well, you can knock it... it isn't... it isn't... I mean this is note form you don't use numbers in an...*

Rater A: *(interrupts) it's not note form*

Rater D: *it is... you don't use numbers in a connected writing exercise...*

Rater A: *I do all the time*

The answer is an example of a phenomenon familiar to many who have taught EAP (English for Academic Purposes): the "pre-sessional overlay". Having been introduced to conventional English academic discourse and exhorted as to its importance in the university setting they are entering, learners at the early stages of assimilating it often adopt the surface

features of the genre without having internalised the underlying logical structure the surface markers normally carry, resulting in a veneer of conventionality over discourse which is deviant in the genre.

Raters A and D disagree quite strongly about whether this veneer of English academic rhetoric (if numbering and laying out points on separate lines may be so named) has any value. Note that they do not disagree about whether the underlying appropriate rhetorical structure exists. they agree that it doesn't. But rater D suddenly reveals himself as a purist: tolerant of alternative rhetorical structures in other essays, he can't accept this 'playing the system'. Rater A takes a more pragmatic approach: "it's a way of doing it". This disagreement expressed itself in a markedly lower score from rater D than from rater A. Rater D assigned a band 3, raters B and C band 4, and rater A band 5.

Rater D had already shown that he objected to a superficial veneer of rhetori al convention over a discourse structure which did not have the internal logi which such conventions normally convey, in the discussion of writer number 11's answer (Extract 5.2.10). We have already seen that raters do expect conventional rhetorical structure and comment on its absence (ee also the discussion of extracts 5.2.6 and 5.2.7): it would appear from the limited data here that they want it more than skin-deep

2.6.3. Length

While length wa clearly another variable which was strongly influencing raters during the e discussions, the problem of what constitutes an "adequate sample" was settled as a result of them. In the third version scoring proçedure raters are told that a length of 60 words is a minimum, and that quoted or plagiarised material from the input text should not be counted in the candidate's word total.

2.7 'SAP' raters in action?

The extracts in this section, and the full transcripts in Appendix F, reveal little to suggest that when raters are rating M2Q1 they see themselves as SAP raters. There is nothing to suggest that what they look for varies across Modules even at the content level in any but the most unavoidable way. The discussions suggest that raters are applying criteria of academic discourse of a general nature, i e., that M2Q1 as rated is a GAP writing test. These criteria were set up for them by the scoring procedure, it is true, but they seem to apply them with a naturalness and ease which suggests they have construct validity in terms of what raters who are language specialists do when rating essays on topics from a range of subject areas. The problem is that this is clearly not what subject specialist faculty do when rating written work by students in their courses. In this regard language specialist raters resemble the undergraduate English faculty in Bridgeman and Carlson's (1983) study, who placed evaluation criteria for written work in the following order of importance:

1. *paper organization*
2. *development of ideas*
3. *paragraph organization*
4. { *addresses topic*
   { *overall writing*
6. *sentence structure*
7. { *appropriate to audience*
   { *assignment requirements*
9. *quality of content*
10. *vocabulary usage*
11. *punctuation/spelling*
12. *vocabulary size*

In contrast, placement of evaluation criteria by a combination of faculty from departments of civil engineering, electrical engineering, psychology, chemistry and computer science was in this, very different, order:

1. *quality of content*
2. *assignment requirements*
3. *addresses topic*
4. *development of ideas*
5. *paper organization*
6. { *overall writing*
   { *vocabulary usage*
8. { *paragraph organization*
   { *sentence structure*
10. *punctuation/spelling*
11. *vocabulary size*

(there was wide disagreement among faculty across departments about the importance of appropriateness to audience, psychology ranking it with paper organization, electrical engineering with overall writing and vocabulary usage, and chemical engineering, chemistry and computer science with paragraph organization and sentence structure).

Content is placed ninth by English faculty and first by faculty in other subjects: our raters in this ethnographic study are behaving like the English faculty, and are clearly not 'SAP' raters as far as content is concerned. The importance given to vocabulary usage also differs considerably, and once again this study suggests that the raters here are behaving more like the English faculty than like the faculty in other subjects.

Faculty in other subjects consistently placed criteria of content and relevance (assignment requirements; addresses topic) ahead of paper organization and development of ideas, while for English faculty these were the primary criteria. The ethnographic study shows that the raters refer frequently and early to qualities of organization and argumentation (development of ideas) in evaluating the M2Q1 essays, and that questions of relevance/irrelevance are the most problematic in making score decisions. Here also what raters do seem to be more like Bridgeman and Carlson's English faculty than like faculty in other subjects. The only criteria on which the two types of faculty agree are punctuation/spelling

and vocabulary size, ranked last by both groups: the raters in the ethnographic study appear to concur in this.

It would appear that this in-depth study of how raters are rating and what they are valuing has been unable to identify any discipline-specific features or values in their decision-making processes. We have not observed SAP raters in action, but rather, GAP raters in action.

CHAPTER SIX

TASK VARIABLES AND WRITER VARIABLES

It was found in the empirical investigation that although the two SAP tests were significantly correlated for the group as a whole and for all Modules except Medicine, in four of five Modules they failed to meet the equivalence criterion of r ≤.80. Also, in four of five Modules the two SAP tests did not share more variance with each other than with the GAP test.

We shall in this chapter attempt to understand the task variables which are operating in the M2Q1, SAPQ and GAPQ questions, and which may be preventing the three writing tests from maintaining a stable set of relationships. There are in this chapter two aspects to the continuing search for 'SAPness' in a writing test: (1) What is a SAP writing test task, and how shall we recognise and replicate it? (2) What is a SAP response from a writer and how shall we recognise it? We will approach this study through a close reading of writers' responses to the three tests, seeking indications from within answers of what it is that writers are responding to in each task. We will also attempt to understand something about the writers and the kinds of knowledge, skills and expectations they bring to these tests The two aspects are interwoven through the chapter so that the essential chemistry of the relationship between the task and the writer's response can be continually stressed.

The chapter closes with an attempt to establish a system of task analysis which can characterize task variables and their relationships, identifying the task variables which contribute most to the SAPness or GAPness of a task, and the task variables which contribute most to close relationships between tasks.

1. THE WRITING TEST AS A COMMUNICATIVE ACT

Houghton (1984) characterises the formal academic essay as "a kind of game in which the writer, according to the extent to which he or she is familiar with the rules and is able to use them, seeks to satisfy the demands of the reader/marker." (p. 47). This is equally true of the formal academic e say test, which is a game with rules at the linguistic, rhetorical, discoursal and pragmatic levels. Ostensibly the purpose of this game, or communicative act, is to convey content. this may be the actual purpose in the case of the true content area essay, in which a faculty member requires an apprentice member of that disciplinary community to display the extent to which she has acquired control of the sh red knowledge of the community. When the actual purpose is to display the extent to which the writer has acquired control of the tools of communication within the disciplinary culture without having an authentic me-sage to convey, there is a real disjunction between apparent and true purpo e. A question which we shall ask ourselves throughout this chapter is whether the disjunction is exacerbated or minimised by the use of SAP tasks as opposed to GAP tasks

1.1. The writing test as discourse

A writing test is, by its nature, a discontinuous discourse. In discourse analytic terms (Coulthard & Ashby, 1975) we can think of any writing test as a discourse exchange where the expected sequence is.

Figure 6.1.1

INITIATION (= the essay question)

↓

RESPONSE (= the student's essay)

↓

FOLLOW UP (= the assessing of the essay)

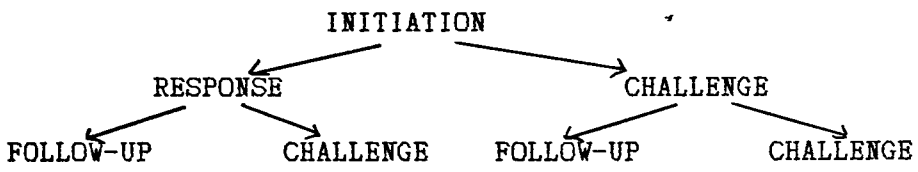
An important difference between SAP/GAP (ESP/EAP) writing tests and tests in academic subjects which require essay answers is that in the SAP/GAP writing test each 'move' in the discourse exchange is made by a different person, that is, it is a three-way exchange, whereas the subject specialist essay test is more likely to be a two-way exchange, as shown in Figure 6.1.2.:

Figure 6.1.2

	SAP/GAP	Subject Specialism
INITIATION	test constructor (P1)	academic supervisor(P1)
↓		
RESPONSE	student (P2)	student (P2)
↓		
FOLLOW-UP	language tutor (P3)	academic supervisor (P1)

Sinclair (1983) adds to the discourse exchange a move he calls the challenge, a term used by Labov and Fanshel (1977) in describing a discourse function in sociological, psychological terms. The addition of this move to the discourse exchange significantly expands the discourse structure (Figure 6.1.3):

Figure 6.1.3



The question answerer, as a participant in a discourse, albeit a discontinuous one, always has two options open: the predicted, or unmarked response ('response'), or the unpredicted, marked response ('challenge'). Challenges may be unconscious (that is, the writer may be unaware that she has replaced the response expected by the initiator with an unexpected response); or they may be conscious. As Figure 6.1.3.

shows, a challenge which replaces a response can be either followed up as if it were an unmarked response, or it can be challenged in turn. This set of possibilities is discussed in sub-section 2.2.

1 2. The writing test as a task in/for an academic setting

As we saw in Chapter 2, Section 3.4, there has been to date surprisingly little research into writing test tasks and rubrics, and the variables within and among them. Surprisingly, because in all writing tests the task is a formal statement of what the student is required to do. If it is negotiable, it is only negotiable within strict parameters. The writing sample produced by the student will therefore be profoundly affected by the task. Ruth and Murphy (forthcoming) say:

> *Writing tasks themselves have rarely been treated as objects o inquiry; thus, we have almost no literature on the ways writin tasks function as instruments of inquiry in either assessmen or research. (ms. p. 420)*

We have only a small number of context-bound studies (e g., Chaplen, 1970; Jordan & Mackay, 1973, Kroll, 1979; Johns, 1981; Horowitz, 1986), and the studies of a larger number of institutions by Weir (1983) in Britain and Bridgeman and Carlson (1983) in the United States to inform the design of writing tests in academic settings. While there was a detailed needs analysis behind G1, G2 and M1, M2 was designed without any needs analysis specifically focussed on what students actually need to do when writing in academic contexts. Bridgeman and Carlson (op cit) asked faculty across university departments to rate types of task in terms of topic type, and found no single topic type that was universally approved. Tasks requiring testees to describe and interpret a chart or graph were preferred by more departments than any other, but were not favoured by undergraduate English departments or MBA programmes. Undergraduate English departments preferred topics requiring testees to compare or contrast, and take a position, but this topic type was not favoured by engineering and chemistry departments. The Bridgeman and Carlson

analysis does not include a detailed task analysis, nor do they propose such a procedure. Weir's (op cit) detailed observations in a range of academic settings revealed too little consistency across settings to enable the observational data to be used as the basis for the design of discipline-based writing test tasks.

This researcher carried out a small survey of faculty at the University of Edinburgh, briefly discussed in Chapter 5, Section 1, to investigate the design parameters faculty have in mind when preparing tests in writing in their own discipline, and the criteria they use when scoring the writing produced by the postgraduate writers on their courses.

48 faculty members, across all disciplinary areas, all of whom had a good deal of experience teaching overseas postgraduate students in their discipline, were invited to a respond to a questionnaire (reproduced as Appendix H). Responses were received from 24, i.e., 50% The results of the survey are discussed in detail in Hamp-Lyons (forthcoming), but are summarized here.

First, the results show that even in maths and 'non-language' courses some writing is necessary: only one of the 24 respondents said writing is never used in the whole course. The results also show that all modes of writing are required with some frequency; and that 50% of respondents use writing for 75% or more of their examination requirement, while 20% said 33% or less of the examination requirement involves writing.

Asked about the design factors they consider when preparing tasks using writing, faculty produced 23 different substantive responses. Fortunately, however, there were multiple responses and thus some instances of agreement among faculty. The frequently cited criterion, with 13 variously worded responses, was the avoidance of ambiguity. Other responses included 'questions which are realistic in the time available ' (5); 'letting students show the ability to apply what they

have learned' (4); and 'covering a balance of the course' (3). There were 47 responses altogether, suggesting that the average faculty member has only 2 or fewer substantive design criteria for tests in writing.

Faculty descriptions of their scoring criteria emphasize the role played by content in score decisions. Because the items were open ended, it was not possible to place the criteria listed by faculty in a priority order, but 49 references relating to 'content' occur, as do 29 occurrences of 'rhetorical' criteria, as detailed in Chapter 5, Section 1. The only other criteria stated by faculty were 'linguistic' (total of 6 responses); 'stylistic' (total of 3) and 'subject specific' (e.g., "correct solution method" (1); quantitative argument" (1); originality (1: Literature): total of 5).

This small study suggests that, while faculty typically do not consciously articulate their design and scoring criteria, they are able to do so to at least some extent when called upon. Further, the responses when totalled reveal a fair measure of agreement as to what is important in design - disambiguation - and in scoring - content - even if faculty are not well able to describe what these criteria look like.

In this chapter, when considering the two SAP writing tests for construct validity and indeed for operational validity, these two requirements - clarity and focus on content - will be borne in mind.

2. WRITERS' RESPONSES REVEAL TASK VARIABLES

In the investigation of task variables through writers' responses which follows, a 'task' is defined as the total input available to the writer. A task consists of: (1) the instructions as to what is being valued, time available, etc., known as the 'rubric'; (2) the materials upon which an answer is to be based, if any (written text, picture, non-linear text,

taped material, etc.) referred to as the 'resources'; (3) the statement of the "agenda" for an answer, known as the 'question'.

We saw in Chapter 2, Section 3 that our present level of understanding of task variables is very limited. Although some parameters for examining task variables have been suggested, there are some major differences between parameters suggested by different investigators. There have been too few studies to permit any confident prediction of the relative effects of different task variables, or of manipulation of variables, and none of these have taken place in contexts similar to the comparison of SAP and GAP writing tests which is the concern of this study.

Therefore, in the investigation which follows, the movement is from what writers actually did in their responses, and towards a characterization of the features of the task which caused them to do what they did. The observation of writers as a method of describing task variables entails attention to 'marked' responses, and two categories of marked responses will be studied.

2.1. Incompetence

In pragmatic terms a writing test is a highly restricted social act within a closed system of knowledge and beliefs, and an academic writing test, even more so a specific academic writing test, functions within an even narrower system of knowledge and beliefs. However, because of the discontinuity of the discourse it is not possible to know whether the same set of knowledge and beliefs is shared by all participants in the discourse exchange.

Many overseas applicants to British universities and universities in other English-speaking countries have very limited exposure to the academic culture shared by these countries and are quite unaware of the social context and expectations within which their writing tests are set.

Many of them are accustomed to an educational system where testing in academic settings is carried out entirely through multiple choice tests or through oral examination. They may lack any test-taking skills for a written test, and are therefore handicapped in the pragmatic component of their linguistic ability. They may lack the necessary competence to take note of the messages sent by the elements of a writing test task. Further, it may be that the setters of writing test questions are themselves not conscious of the messages they are sending out, or of the fact that they have failed to convey exactly their intended 'message'.

Investigation of testee responses in the database used for the empirical investigation in Chapter 4 revealed occurrences of what is referred to here as 'incompetence' (in the non-pejorative sense of 'lack of competence'), relating to each of the three elements of the writing test task. These are: misinterpretation of the question; misinterpretation or misuse of the resources; misinterpretation of the rubric. There are, in addition, occurrences of general pragmatic incompetence which relate to underlying assumptions and values brought to the test by the initiator and the assessor but not shared by the student respondent.

2.1.1. Misinterpretation of the question

Misunderstanding of the question appears to arise mainly as a result of failure to comprehend specific vocabulary items in the question. In the GAPQ question, for example, the word 'sin' is central; a number of testees clearly did not understand the word, or understood it only vaguely. Some of them deduced the meaning from the context, but deduced incorrectly, as did writer number 121 (Extract 6.2.1.):

Extract 6.2.1.: Writer No. 121

> *I think that there is good and bad in every one is the most serious sin mentioned in the text. I think that is true because we can see some body one day is very nice and very*

friendly person but one day you might find him in a very bad hapit. The person himself is some time devil.

The kind of vocabulary items which result in misunderstanding of the question are not only content/concept vocabulary, but also 'strategic' vocabulary, as shown in the following extract from writer number 42's answer to the SAPQ (SS) question (Extract 6.2.2.) in which the writer, asked to describe the effects of the fall in death rates describes instead the causes.

Extract 6.2.2.: Writer No. 42

It is seen that in Britain and Western Europe, the expectation of life is about 70 years. This is mainly due to the development in medical knowledge and sanitation system. A second factor involved is the benefit of advanced agriculture and industry which gave a better standard of living. ...At the end it is concluded that application of scientific and medical advances has resulted in the death control in Western Europe.

Misinterpretation of the question appears to occur unevenly across the questions in this database. GAPQ was quite often misinterpreted, usually due to unfamiliarity with the vocabulary item 'sin'. The M2Q1 LS question was often misinterpreted: the input text is about the 'green revolution', and the question opens with a reference to the green revolution, but the core of the question asks testees to write about "modern farming methods". Many writers did not take note of the distinction, and wrote about the green revolution (see, for example, M2Q1 No.15 in Appendix G). None of the other questions appeared to generate consistent misinterpretations.

2.1.2. Misinterpretation or misuse of the resources

Resources are provided to give testees something to write about and to support them in their structuring of an answer. It can happen, however, that the resources are themselves misunderstood and instead of being a

support are in fact a cause of an irrelevant or untrue answer. Writer number 53 responded to the GAPQ question with reference to 'adultery', which appeared in the input text in the following context: "Britons rated the prohibitions on adultery and coveting thy neighbour's wife higher than did any other nation." This writer, however, who in the SAP questions took the Life Sciences Module, interpreted 'adultery' as 'adulteration' (Extract 6.2.3.):

Extract 6.2.3.: Writer No. 53

> *... Adultery has a bad effect in population. For example if any body adulterated any food it can cause serious illness, can produce disease, even can cause death of many people at a time. ...Not only food, adulteration can also done in chemicals, building materials, cosmetics and other useable commodities. ...So, it should now be clear to everyone that a sin like adultering has got a worst effect on population.*

Writer number 80 found the SAPQ test input text for the GA Module too long and difficult, judging by the fact that she never became aware of the text's movement from apparent good effects to actual bad effects of the 'green revolution' in India (Extract 6.2.4.):

Extract 6.2.4.: Writer No. 74

> *The idea of the green revolution was embraced enthusiastically by the New Delhi government and the number of the number of the hungry Indians was increasing remorselessly.*
> *The 1965 Indian 5 year plan swung alot of money from the government. And in 1970/71 green revolution reached its highwater in India and it helped alot of people to eat and to not be hangry becuse it is cheap and the have less money to buy it*

Most of what this writer produces is plagiarised from the text, and from the early part of the text. The student does not demonstrate understanding or even reading of the later part of the input text.

Plagiarism occurs frequently in answers and is the main misuse of resources. Many cultures not only do not share the strict anti-

plagiarism stance of western academic culture, but may even value the selective use of the exact words of respected authorities. The provision of text as a resource may suggest that such use of the material in it is acceptable, despite the rubric instruction to the contrary. We shall examine the relative frequency of occurrence of plagiarism in sub-section 2.1.4.

The resources of certain tasks seemed to generate more misinterpretation than others: there were frequent instances of writers unable to "read" a bibliography, so that they did not know whether a source was a book or an article, which was the author's first or family name, etc. But more significant misinterpretations occurred in the case of M2Q1 ME, where many writers failed to understand the relationships drawn in the input text between the experience a trainee doctor can get from working as a paramedic and the benefit this experience offers to them as doctors in training, presenting a very anecdotal response rather than the reasoned one which the question/resources combination made possible and preferable (see, for example, M2Q1 No.116 in Appendix I). Probably because of the misinterpretation of the term 'sin' in the question, GAPQ generated a number of misinterpretations of resources. It seemed that writers, as readers, approached the input text with a preconception of what 'sin' was and what data they were looking for, and found enough references in the text that fit their expectations so that they did not need to go back and reinterpret the question. Writers seemed to find the resources for SAPQ SS difficult to interpret, probably because the input text was quite long and not a great deal of it was relevant to an answer to the question: as we shall see below, SAPQ SS generated short answers, and part of the explanation for that may be found here. Although all resources were misinterpreted by some writers no others suffered a consistent, significant misinterpretation.

2.1.3. Misinterpretation of the rubric

The most significant misinterpretation of the rubric on any of these three tests occurred on M2Q1, where students failed to take note of the information that they should spend 25 of their total 40 minutes on Q1, and instead spent more time on Q2 than Q1. Because of the scoring system used for M2 from 1980 to 1985, this resulted in very low scores for those students. This is not something which can be accounted for in interpreting the data, but it does suggest a note of caution for future test design.

Another type of incompetence relating to the rubric was plagiarism, which occurred on all three tests but was least noticeable on GAPQ. The culturally different interpretations of the meaning of plagiarism, and of the value of the 'expert' statement as opposed to a student reinterpretation and rephrasing, seem to be so powerful that it overrides the writer's reading of the rubric. For many writers, the existence of an input text and the direct reference to it ("Refer to pp x - x in your Source Booklet.") is construed as an invitation to repeat the ideas and the language of that text. It would appear that a rubric which says this, and also says "If you use information from the Source Booklet, put it in your own words.", is sending a conflicting message, the exact interpretation of which is outside the present pragmatic competence of many student writers.

A problem occurred on GAPQ, which was not exactly a rubric misinterpretation, but rather one of test layout, which combined with writing test incompetence to lead some writers astray. GAPQ was set out with the rubric at the top of the paper, then the input text, then the question (Appendix A3), leaving blank about 25% of the front of the sheet. The rubric states: "Read the text below and then answer the question under it." The meaning here is 'the question which is under it', but a number of writers interpreted this as meaning 'the answer must be under

it' and wrote on the front, and only the front, of the sheet. Although an estimation of the relative length of answers does not suggest that GAPQ answers are shorter in general then M2Q1 answers, some of which were very short, and while the reading of these answers does not suggest that they are left incomplete, the possibility that this layout flaw affected scores remains.

2.1.4. Pragmatic incompetence

'Pragmatic incompetence' refers to instances where writers show themselves to be unaware of some fundamental behaviours expected of writers in western academic culture, lacking a certain "test wiseness" that all students educated within the culture acquire, certainly by the postgraduate level upon which this investigation is focussed. Some of these fundamental expectations are overtly marked in the test rubric: others are not.

2.1.4.1. Plagiarism

We have discussed plagiarism both as a misuse of the resources and as a misinterpretation of the rubric. Plagiarism is a major category of pragmatic incompetence, where initiator and assessor share a set of cultural values that are not shared by many student writers from other cultures. Discussion of the transcript of the piloting of the M2 Assessment Guide in Chapter 5 revealed cases where writers had plagiarised, and where raters reacted more or less strongly to their doing so. While raters with wide experience of the resource-based writing of non-English speaking students learn to be more understanding of plagiarism, the problem it creates is that the rater does not know how to value what the student has written either as content or as rhetoric.

Occurrences of plagiarism in this database suggest that it is more likely to occur when the question is closely related to the text structure of the

input text, or when the writer is personally unfamiliar with the subject matter, or when she is unengaged by it. Thus GAPQ generated few occurrences of plagiarism other than of isolated phrases because the question did not follow the text structure; M2Q1 GA/SS also generated little plagiarism, since the input was not a text in the usual sense but an annotated bibliography. M2Q1 LS, on the other hand, generated quite a lot of plagiarism, as did M2Q1 ME. M2Q1 PS was related to a detailed table rather than a linear text, and writers were able to use much of the terminology from the table, although there was no continuous prose they could plagiarise. As we saw in Chapter 5, raters were suspicious as to whether writers were being helped by the amount of useable vocabulary and structure provided by the resources, and tended to rate more harshly. On SAPQ, the GA question seemed to encourage plagiarism but of appropriate material, while the LS question seemed to encourage inappropriate plagiarism (it appears to be a convention of academic discourse that plagiarism, once recognised, is penalised equally whether or not it has been well done). The SS question appeared to allow both appropriate and inappropriate plagiarism, because there was so much material in the input text, some relevant and some not. The ME question, like M2Q1 GA/SS referred to a bibliography and so plagiarism was not possible: the PS question was rather minimally related to the input text and again plagiarism was not an option.

2.1.4.2. Length

We saw in Chapter 5 a discussion by raters of what constitutes an 'adequate sample', and we learned of the decision that this would be defined as at least 60 words. In the version of M2Q1 used in this database, however, the rubric states "...write 15 to 20 lines". Some writers wrote much less than this, or much more. Those who wrote less rarely received scores which would indicate competence (i.e., 5 or above): others wrote much more and typically received much higher scores. We saw in Chapter 2,Section 3 that length is very often found to be a

significant factor in explaining scores, and although no statistical study has been made of the effect of length in this corpus, the overall impression of reading answers over and over and looking at their scores is that something similar is happening. To write less than the minimum stated by the rubric is pragmatically incompetent.

Examination of answers suggests that GAPQ generates a wide range of answer lengths. M2Q1 GA generates rather short answers: however, when the same question is answered by SS students the answers seem to be much longer. M2Q1 LS and ME generate long answers while PS generates answers of moderate length. SAPQ GA, which is on the same topic as M2Q1 LS but a different text, also generated long answers; LS and ME generated long answers while SS generated short answers and PS generated amswers of moderate length.

2.1.4.3. Covert expectations

Both plagiarism and length are overt expectations: requirements were clearly stated in the rubric. In M2Q1, and in GAPQ and SAPQ due to their design basis in the parameters of M2Q1, however, certain key expectations of an academic/ specific academic writing test task remain unstated.

No audience is stated for any of the tests: it is assumed that writers have some sense of the context of the test and create an audience for themselves. Although it has been popular to assume that it is necessary to state an audience for a writing test because of the evidence that writers write at different levels according to the audience specification, we are beginning to understand that careful audience specification implies no other readership for the text; it implies the assumption that the writer perceives her relationship to the specificed audience in the same way as did the test constructor; it implies that the test constructor has a clear basis for a decision about what audience to specify. Because writers' responses <u>are</u> influenced by audience

designation, it is beginning to be suggested (for example, by Ong, 1975; Elbow, 1986; Park 1986) that to specify the writer's audience for her imposes an unrealistic constraint. Park (op cit) suggests that it is more authentic to set up the task as a task within a genre, and let the writer's sense of the genre determine the audience. This seems to be what happens with M2Q1, and by derivation with SAPQ and GAPQ. Close reading of the responses in this database does not reveal any instances of pragmatic incompetence in this area: all the writers show an appropriate sense of audience, within their general competence with the written code (that is, the few writers who are at a very low level of control of the written code do not show a sense of audience at all).

Another relatively covert expectation of the writing on these tests is the mode of discourse. Mode of discourse is indicated only by the command verb or verbs in the question (Explain; Discuss; etc.) and the rubric does not attempt to make clear the underlying assumption that the writer will write expository discourse. The SAPQ questions are all purely expository in mode: GAPQ is argumentative; M2Q1 SS/GA, LS and TECH are expository, ME is expository but with some room for argumentation; and PS is argumentative. All the M2Q1 questions ask for some description, narration or opinion from the writer's experience, and GAPQ asks for the writer's opinion. None of the SAPQ questions require the writer to refer to herself at all. Exposition and argument are the classic modes of discourse for the genre of academic essay tests, and all of the writers who produce an adequate sample show themselves to have some sense of awareness of the mode of discourse they should be writing in, although they may not have mastered it. Many writers did not, however, respond to the requirement of the inclusion of some element of personal writing in the primarily expository tasks: this is most noticeable with M2Q1 LS, which asks the writer to draw on her own experience in discussing the advantages and disadvantages of modern farming methods. None of the writers followed this requirement (and there was no indication in the study of raters that writers were penalised for not doing so).

Although there were too few cases of Technology to retain them in the database, the writers who did write on this question made only marginal reference to the advantages and disadvantages of particular metals "for a purpose with which you are familiar". In contrast, writers seemed to have no problem with the mode of discourse of GAPQ, which asked them to make choices between a number of possible 'sins' and justify their choice: it was, then, essentially a personal response argument, another familiar mode of discourse to most student writers. It seemed that the combination of exposition and personal response created an unfamiliar mode of discourse for many writers, i.e., they were pragmatically incompetent in that mode although not necessarily in others.

Audience and mode of discourse are each strongly related to the purpose of the task, and it seemed that none of the writers were in any doubt as to the purpose of the writing tests. Apparently it was not necessary to specify any of these parameters for these writers: all who were minimally competent linguistically were pragmatically competent to this extent.

We saw in Chapter 5 that raters were very concerned about the rhetorical structure of answers. Rhetorical competence is another aspect of pragmatic competence, but it is a much larger set of possibilities than are audience and mode of discourse, and the writers in this database demonstrate great variability in this regard. We looked at some contributing factors in Section 2.6. of Chapter 5, and in the next section we shall consider the relationship between the text structure and the required rhetorical structure of an answer in permitting a writer to demonstrate the best, or otherwise, of her rhetorical competence.

2.1.5. Incompetence - writer or task?

The category of 'incompetence' and its subdivisions emerged from the researcher's close contact with the answers over a long period, and began as the observation of a small number of strongly marked cases of each

kind of incompetence. Once the database was consciously studied for further evidence and instances of these types of incompetence the applicability of the categorisation, for this corpus, was confirmed. But this continual re-reading of answers for evidence of certain variables revealed that instances are not uniformly distributed, across tests, or across questions, or within writers. The question then arises, whether 'incompetence' is located within the writer, within the task, or somewhere between the two.

2.2. Challenge

We defined a 'challenge' as a marked or unpredicted response, and must now refine that definition by limiting it to occasions where misinterpretation of the task, at least at the linguistic level, does not occur, since such occurrences have been classified here as incompetence, and are discussed above. Challenges may be unconscious (that is, the writer may be unaware that she has replaced the task with a task more to her taste) and as such will probably not be marked on the surface structure of the text; only by directly comparing task and answer will the mismatch be seen. Or challenges may be conscious but covert: although the writer is aware that she is replacing the task with one of her own she makes no reference to the replacement on the surface of the answer. This type of challenge cannot be distinguished from the unconscious challenge except through a more sophisticated research technique such as grounded ethnography, which was not used in this data collection: a claim for its existence is at this stage based solely on haphazard anecdotal evidence. Overt conscious challenges are challenges which are clearly identifiable because the surface of the text is marked by an intrusive verbalisation of the challenge, and it is these which we shall primarily use in our search for task variables through writers' responses.

Just as in oral discourse every challenge is an interruption of the flow of the normal social conventions of talk, in written discourse every

challenge is an interruption of the pattern of the normal social conventions of writing for a reader. Because in academic and specific academic writing these conventions are stronger than in casual written modes such as letters and lists, challenges are more noticeable. Because they do not conform to the discourse expectations, challenges may seem pragmatically incompetent. The incidence of challenges in this database, however, suggests that challenges are made primarily by pragmatically competent writers and are actually a manifestation of pragmatic competence. When challenges are made by less pragmatically competent writers they appear to arise from a very strong emotional response.

When a response is replaced by a challenge, the follow-up may as a result be replaced by a further challenge: that is, in the case of the writing test, the rater may choose to treat the writer's challenge as a valid response and score it in the same way as any other response, or he may choose to treat it as invalid, and score the answer lower because he sees it as invalid. Just as the writer's challenge may be unconscious, conscious but covert or conscious and overt, the rater's challenge may be any of these.

In most cases of challenges in this corpus, it is not possible to know whether raters reacted to answers with a follow-up or with a challenge of their own. There are, however, two clear cases of challenges in the smaller corpus used for the ethnogrpahic study of raters reported in Chapter 5, and these discussions will be reported with the writers' answers in the following discussion.

2.2.1. Why do writers challenge?

Weaver (1983) studied teacher candidate writers in the process of composing in response to a range of tasks, and found that they needed to go through a process which involved attending to the task and assessing its value to them. If they found it of value they "transformed" to a

self-initiated one (in our terms, they made an unmarked response); if they did not, they "replaced" it with a task they could value (in our terms, they made a marked response or challenge). The study of challenge data here is based on the claim that writers only challenge when they are disturbed by a task in some way, to the extent that they are unable to value it as it stands sufficiently to make the necessary transformation, to take it to themselves. Marked responses of this kind are "noise" in the system, that is, they are indicators of systemic malfunction, unlike the various types of incompetence which are evidence of writers' skills in specific areas in their own right. We shall look here at frequency of occurrence of challenges on different questions and the (apparent) grounds of the challenges, and in the next section we shall consider how we can apply what we have learned from the investigation of grounds of incompetence, and of challenge data, to devising a system for identifying and describing task variables and their impact in writing test tasks, especially specific academic writing test tasks.

2.2.2. Challenges to GAPQ

GAPQ generated few challenges: the most overt challenge was from writer number 52 (Extract 6.2.5):

Extract 6.2.5.: Writer No. 52

> *I think it is very difficult to answer this question. The reason is that, as ethics is a subjective concept in each particular community, the meaning of the "sins" is varying a lot.*
> *Furthermore ethics refers not only to the community but also to each one of us seperately. So, the action of committing a crime has a different meaning to each one of us and to each one community. Subsequently we have to judge each case according to the customs, that is according to the way of life of each particular community. For example, it is generally believed that killing is the most terrible "sin". The people supporting it think that we can not do something like that because life is given by the God and we have not the right to kill someone irrespective of the reason. But these people can stand*

the everyday starving of thousand of people of the Third World while they know that it is essentially a way of killing all these people by the developed countries.
(etc.)

The aggregate score for this writer on GAPQ was 5. The question arises, in relation to this answer, of what made this piece of writing seem to the readers that the most positive thing to be said about it was "the basic message is presented", and of what made the readers feel that the "logical presentation (may be) broken and lack clarity and consistency". The thesis of the answer is very clear, and it moves from general to specific in conventional ways, arguing the thesis through supporting examples in ways which are usually valued in western academic culture. We cannot attempt an answer at this stage.

The next example of a challenge to GAPQ is more subtle: writer number 66 does not tell us she is making a challenge, but leaves us to gradually reach that conclusion (Extract 6.2.6.):

Extract No. 6.2.6.: Writer No. 66

1. *A man's belief in sin is one of the most serious sins. because when he does somethings, he immediately thinks whether or not there is any good effect out of his works. If he feels that there are some evil effects out of his works, it is his sin which leads moral depression upon him.*
2. *Another serious sin is the regretness of people. A man having done something should not regret that he is wrong. If he always thinks that what he does is wrong, it keeps him unhappy most of the time. That's why the rich people are forced to be unhappy in comparison to the poor.*
3. *Another serious sin is the believe in God. Because nobody can give any proper definition about God. But some people are regular-worshipers. They have faith in God and consequently thinking about God they sometimes show their strict attitude towards others. Sometimes such strict attitudes brings potential harmfulness to them.*
(etc.)

This writer also scored an aggregate band 5 for this answer. Since at the level of linguistic appropriacy and accuracy the answer seems to be less competent than that in Extract 6.2.5., the readers presumably valued some other aspect of the answer a little more in this case. Once again, however, we may wish to ask ourselves what characteristics of the answer led raters to find "the basic message is presented" the most positive thing they could say about it.

2.2.3. Challenges to M2Q1

It is very striking in this database that some M2Q1 questions generated many more challenges due to inability to value the task than did others. Although M2Q1 PS had only 7 testees, there were 5 such challenges. Extract 6.2.7. shows how writer number 31 makes a strong overt challenge :

Extract 6.2.7 (a): Writer No. 31

> *Even though I am a scientist, I strongly consider the opposition to new scientific processes as a healthy action. I think that man's new experiments should compromise their experiences and the environment, in order to avoid damages to the present and future nature. Basically, no man or nation has the right to, in the name of a scientific progress, destroy their own habitat - the only one we have now.*

The reader may feel unsure whether this is in fact a challenge, or an instance of incompetence due to careless reading of the question (reading "oppose" for "defend"). But as the answer continues it becomes clear that this is not the case:

Extract 6.2.7(b): Writer No. 31

> *Fortunately though, I believe that man can find better ways in order to garantee the environment and mankind preserved. Indeed, I also believe that new scientific processes can be done just to improve the quality of life on Earth, in despite of all economic interests involving this man's action all around the world. Even though*

someone can find this thesis completely utopic, I really trust it.

The writer is clearly aware that she is running counter to the viewpoint she has been asked to argue, and that "someone" might not like that, but she is prepared to stand by her belief. Writer number 31 in the main corpus was writer number 19 in the ethnographic study, and examination of the transcript suggests that the challenge is treated as valid (Extract 6.2.7(c)):

Extract 6.2.7(c): Transcript of Writer No. 19/31

Rater A: *I should... because there are no or... I can't see any... linguistic inaccuracies, come up with a 9. I'm very loathe to do this... so in fact I've given it an 8. Now I'm not sure that my reasons for not giving it a 9 are very good ones but anyway that's my... I see it as a challenge to the question (murmurs of agreement)... er... quite clearly... um... and I don't... you know... I think we've agreed... or at least my feeling is you can't... er... mark that down... um... because it's very well argued (murmurs of agreement)*

This was the first turn in the discussion, and the fact that the answer is a challenge is not referred to again. Although we cannot know what goes on in raters' minds, there is no evidence from what they say that the recognition of a challenge rather than a response caused them to lower their scores. They appear to accept the challenge as valid and respond to it with a follow-up.

The situation is rather different with the answer of writer number 5 in the ethnographic study (writer number 120 in the main corpus), however. Here is the answer of writer number 5/120 (Extract 6.2.8(a)):

Extract 6.2.8(a): Writer No. 5/120:

> *The continual use of such potentially harmful processes will change the percentages constituents of environments & for this we must try to minimise the side productions such as dust, soot & sulphur compounds which come from factories.*
> *To controle these products we can try to absorbe them by many modern methods to get a great useful by using them to other products as a primary constituents & not let them to harm our environment. And we can also find other methods by which we will not get much side harmful products also we must bild the factories away of centres of towns & to increase cultivations around towns & big cities to decrease dust & other impurities to harm the environment.*

The reactions of the raters are presented in the following extract from their discussion (Extract 6.2.8(b)):

Extract 6.2.8(b): Transcript of Writer No. 5/120

Rater D: *(...) there's a problem of course in that he doesn't answer the question*
Rater B: *I wondered when you'd get to that*
Rater A: *Thank you - precisely*
Rater B: *"As a scientist how would you defend...*
Rater A: *How can you give him any mark at all? He hasn't answered the question.*
(...)
Rater B: *To a certain extent in fact he argues against... the position he's supposed to be taking - and you... you might take that as an indirect challenge to the question but um... I think that's probably unfair*
Rater A: *No, I don't think he's even noticed it...*
Rater D: *Well it's not entirely irrelevant in that it does suggest or begin to suggest methods for reducing (inaud)*
Rater B: *But that's still not the point of the question*
Rater A: *Yes but he doesn't defend it... he's supposed to be defending... the continued use of such potentially harmful processes... on what basis, you know... that you're benfiting mankind or you're...*
Rater D: *it all hangs on "defend" doesn't it? It all hangs on whether you interpret "defend" as being... how would you put up a case for them,*

or how would you... get round them, or minimise them...

Rater A: *But, but it's... no... I don't think... I mean, "defend" is "defend"*

Rater B: *All we needed is a simple statement like 'society needs these... the products of these processes, so all we can really do is... is... accept them and do the following to reduce them'... and er... I mean... I was... it struck me immediately that... as being irrelevant*

We may wonder why the raters can respond so differently to two answers to the same question seen in the same rating session. Rater B's final turn suggests that the raters did not accept this as a challenge because it was not overtly marked on the surface of the text, and that if it had been they might have regarded it with much more favour. It is impossible to know whether this challenge was conscious but covert, or unconscious. Writer 5/120 responds from within her disciplinary knowledge, putting forward methods for handling the problem: writer 19/31, in contrast, responded as a community member, putting forward not arguments or practical suggestions, but polemic based on moral outrage. Was it this to which raters responded favourably, or was it the evident lower linguistic level of writer 5/120's answer to which they responded unfavourably? It is impossible for us to know.

In Extract 6.2.9., writer number 122, responds similarly to the M2Q1 PS question, but the challenge is overt:

Extract 6.2.9.: Writer No. 122

The phisical facts of pollution can be measured by using scientific equipments, and scientists know the process of the facts. Engineers who know the scientific knowledge only can develop facilities which reduces this harmful processes. On the other hand, politicians and exectives of companies have a force to decide the use of the beneficial but harmful processes. The decision must be or would refrect the will of people who are enjoying and are harmed by the process.
Therefore, the scientists only can give people proper information about the process, and the engineers only can

give people proper information about the technology and the cost of preventing the harmful effects.
For me, the question above does not make sence. The choice of continual use of such potentially harmful process or cutting off the use of the process does not depend on the scientist. Scientists want to know everything in a rational way. The knowledge obtained by this way is so repeatable and testable, or reliable, that this knowledge have a power. The way how we use the power is not on the responsibility of the scientists.

While this answer displays clear linguistic weaknesses, they intrude only marginally into our understanding of the argument. The answer is well-organised, and it challenges the question not from personal opinion, but from the position tradionally held by scientists, most of whom leave the moral issues of scientific developments to others in just the way she describes. In other words, this writer is writing from within her discipline and its values. However, raters appeared not to have acknowledged this: the aggregate score for this answer was band 5. As with the answer in Extract 6.2.5, we must wonder why it was that raters found so little of value in the answer, and felt a description of "theme can be followed, but logical presentation may be broken and lack clarity or consistency" was the most appropriate.

The M2Q1 SS/GA task also generated a number of challenges: most of these came from writers in the SS Module. Writer number 87 makes overt a challenge it would be hard, reading the answers carefully, not to notice was covert or unconscious in many answers to this question (Extract 6.2.10.):

Extract 6.2.10.: Writer No. 87

At this moment I am not really interested in prison's problems. Surely, Size's Prisons I have known is a well documented book and I might learn a lot about prisons reading it. I would like to know about problems of "open prison". However, if I have to choose only one book to read, I prefer to read Henry's Who Lie in Gaol. I expect it to be like a novel. I would like to find an easy book to read. Aditionaly, I think I will find interesting to

hear of personal experiences of a prisoner. Generally, one has opportunities to know opinions from the police, authorities, etc, but rarely there is opportunity to know what prisoners think. I feel I have a professional interest in this book. As a psychologist I am interested in knowing emotions, reactions, feelings of a man in such a hard situations.

This answer was scored band 7 ("theme presented in a well-ordered, intelligible manner with well-structured and relevant supporting detail"). The answer (reproduced in full above) consists of a single paragraph although not just a single idea. The first three sentences are irrelevant, and the first is a challenge. There are some fundamental linguistic errors ("rarely there is"), not a great deal of support, and it scarcely seems "well-ordered". It would appear that the challenge and the two sentences about the other book 'counted' in some way toward the judgement made about the answer. A similar situation arises with the next extract.

Writer number 93 also responded to this question as an SS writer, and it would appear that the whole answer is a challenge (Extract 6.2.11.):

Extract 6.2.11.: Writer No. 93

If I want to choose one of these books which are mentioned above, it depend on what and why I want to study either to build up a background or to collect a date to make a research. if I have to choose one, I will choose the book which is written by size's, as he had an experience, and he was dealing with different type of prisons, so any one interested to write something about the real life of prison, this book will provide him with a real information. But that it doesn't mean the resarcher just depend on one reference like this kind. he should have collect his data from different resources to make a good decision.

The writer first appears to object to the lack of a stated purpose, or context, for the task, which leaves her without an authentic basis for a choice. The use of "have to" suggests her sense of being coerced, and of

playing the tester's game, but unwillingly. The expression 'if I had to' is used very frequently as a lead-in to answers to this question. Her second challenge ("that doesn't mean") is not simply to making a choice without knowing what the basis is, but to making a choice at all: it is a challenge from within at least the general academic community, and perhaps from within the values of her particular discipline, where there are no absolute 'proofs' and knowledge grows through a more gradual and consensual accumulation of evidence and understanding. The score of band 6 seems reasonable for this answer.

2.2.4. Challenges to SAPQ

The SAPQ questions generated very few challenges. It is noticeable that SAPQ PS did not generate any challenges, although the topic of the question is very close to that of M2Q1 PS. SAPQ SS and ME did not generate any challenges either, and the closest to a challenge which arose for SAPQ LS were two or three plaintive comments like this one from writer number 52: "Unfortunately the answer is not straightforward. It depends mainly on how many pairs of genes control the particular genetic characteristic of cows." The writer appears to be letting the test constructor know that the task is unreasonably hard. SAPQ GA asked writers to "Explain why the 'green revolution' of high technology in food production has created serious social problems in India." A number of writers retained the structure of the input text, writing too much about the good effects of the 'green revolution' before they began to discuss the bad effects. The question is similar to M2Q1 PS in that it asks writers to put only one point of view, and many did not acknowledge that instruction. The impression was not, however, that such responses were covert challenges but that they resulted from a form of incompetence. The answer by writer number 36 comes closest to a real challenge (Extract 6.2.12.):

Extract 6.2.12.: Writer No. 36

(over)

Actually, the 'green revolution' has brought a great advantage. However, we should pay attention to another aspects of it. The new varieties of seeds need to be pampered or they sicken and die, and have to have regular supllied of water. As a result, just irrigated fields can be planted. This means they need expensive artificial fertilisers.
Moreover, the green revolution caused a change between the rich and poor farmers. In other words, it resulted in a discrimination. Because only well-off farmers can buy the new seed and only they can take a risk in doing so. Anyway the poor farmers still remain poor.
In my opinion, what is called 'green revolution' in India will be called 'Catch 22', which will be always in backfire.

The writer appears to be only prepared to accept a halfway position in terms of the stated point of view of the question, and is at some pains to point out the good effects of the 'green revolution' also.

The variation in the number of challenges from question to question can be clearly linked to variables in the design of the task, as we shall see in the next section.

3. TASK VARIABLES

The study of two kinds of marked responses, incompetence and challenges, has provided us with some data upon which to build a tentative model of task variables. We saw that incompetence can arise from the writer's response to the resources, the rubric or the question, or to the interaction between them. Resources which are too short or too indirectly related to the question are unhelpful, but resources which are too long are a source of difficulty, and resources with a text structure too closely matched to the rhetorical structure implied by the question tempt plagiarism. We saw that expectations concerning plagiarism and length need to be clearly specified and even then writers from some cultures may not take heed of them. The lack of a specified audience did

not appear to cause a problem in this corpus, but we have no comparative data to indicate whether better writing might have resulted from a specification of audience. Mode of discourse was indirectly specified through the verbs 'discuss', explain' etc., and it was not uniform across questions. Since other studies have shown that exposition is more difficult than narrative or description, and argument is more difficult again, we might expect that the argumentative questions would yield lower scores than the others. When two modes are required in one question the task can be expected to be even harder. The requirement in the M2Q1 questions to bring in personal opinion or experience was obeyed for GA/SS and PS but ignored for LS and received only marginal acknowledgement in ME. There were no indications that students found it helpful or that it led to higher scores.

Challenges occurred most often on M2Q1 PS, a question which was both argumentative and personal - the hardest according to the hypothesis above. In this question also the writer is not given space to make up her mind: she has to argue in defense. It seemed to be this pre-emption of the kind of response which would be legitimate which most (5/7) of the PS writers challenged, rather than the difficulty of the question. The challenges to GAPQ seemed to be a rejection of the basic premise of sin as a universal concept - a premise deliberately chosen in an attempt to avoid a culturally biased topic. SAPQ GA also pre-empts the direction of response but perhaps because the text structure leads the reader/writer to the conclusion which is the starting point of the question, the question generated few challenges. M2Q1 GA/SS took a topic unfamiliar to almost all the writers and provided very little in the way of input resources: unsurprisingly it generated quite a lot of challenges.

3.1. Categories for task analysis

In attempting to develop a system of task analysis appropriate to the study of SAP and GAP tasks, the categories proposed by Pollitt et al (1985) have been taken and adapted, because (1) they seem to be the most fully developed and researched to date, and (2) they offer the system which comes closest to describing the phenomena observed in the writers' responses in this corpus.

3.1.1. Subject difficulty

Pollitt et al's first broad category is subject or concept difficulty: as this has been adapted for this study, this category refers to: 1) the writer's degree of familarity with the subject matter; 2) the difficulty level of the resources upon which the question (and therefore the answer) is based; 3) the abstractness of the concepts to be discussed. Subject difficulty does not reside in the subject matter but in the individual and her experience of it, and in the context of a large-scale pre-acceptance test it will never be possible to verify each writer's degree of familiarity with the subject matter before deciding which test to offer her. Although text difficulty level can be established independent of readers, it is difficult to know where to pitch the reading level of the resources for a writing test which is intended to yield meaningful scores at all levels of writing proficiency. Of the three subject difficulty variables, it would appear that only 3) can be absolutely determined, but even to determine the appropriate level of abstractness of the concepts to be discussed entails a number of prior decisions about what writers need to do in real academic or specific academic situations, and such decisions require a detailed needs analysis.

3.1.2. Process difficulty

Pollitt et al's second broad category (op cit) is process difficulty, which here refers to: 1) degree of difficulty of the process 'mode' chosen; 2) strength of relationship between the structure of the task (resources; question) and the potential answer; 3) amount of support in the task (resources; question) for generating appropiate content for an answer. Research was described in Chapter 2 which suggested that some modes of discourse are more difficult than others; narrating, explaining and arguing are not processes at equivalent difficulty levels. While it is commonly assumed that "personal" writing is easier than interactional writing there is little evidence that this is the case, and none that adding personal writing to an interactional task makes it easier. Neither can it be assumed that a close relationship between the structure of a question and the intended structure of an answer makes the task easier: in this corpus such a relationship appeared to tempt writers into plagiarism, despite the warning against this in the rubric. The same is true of the relationship between the content of the resources and the content requirements for the answer.

3.1.3. Question difficulty

The third broad category based on Pollitt et al (op cit), question difficulty, refers in the context of this study to: 1) linguistic difficulty of the sentence or sentences which convey the particular 'stimulus' to which the writer must respond; 2) outcome space, i.e., the space within which legitimate answers can occur. It should be possible to establish more concrete parameters for linguistic difficulty than for other task variables, since the linguistic system has been extensively studied. But to investigate the linguistic difficulty of a writing test task the linguistic structure of the question must be analysed from the point of view of the components of its structure which are significant in

determining writers' responses. There are no reports in the literature of such attempts.

The analysis which follows is this researcher's attempt to establish a fairly simple system of analysis of a writing test question which will explain some of the key linguistic variables influencing writers' responses. In this system four elements in a question are recognised (Figure 6.3.1.):

Figure 6.3.1.: Analysis of Linguistic Structure

	Component	*Description*
1.	*Topic*	*N or NP; assumed to be old information for the writer; open set*
2.	*Comment*	*instructional V or VP and other initiators; closed set each with closed sub-set*
3.	*Focus*	*topic-narrowers; indicate illocutionary force intended for the answer; large but finite set*
4.	*Perspective*	*determines viewpoint to be taken; defines what can be accepted as 'true' by each participant in the discourse exchange*

All of these terms occur in the work of Van Dijk (1977) but are not used in precisely the same ways. The term 'topic' is used here as it is used in discourse studies, closely paralleling the concept of 'aboutness' in philosophy and 'subject' in logic and formal sentence grammar. 'Topic' and 'comment' are a familiar collocation, having the same relationship as 'given-new' in text grammar. Van Dijk (op cit) uses 'comment' to parallel 'predicate' in logic and formal sentence grammar, but for the analysis here 'comment' is restricted to the part(s) of the question structure which tells the writer what rhetorical expectations the answer should

conform to. In cognition, 'focus' denotes attention and is propositionally-based (that is, the focus is not on individual lexical items or concepts but on their function in a specific network of relations). Semantically it is identified with new rather than old information, like comment and unlike topic. Pragmatically it determines relevance, i.e., what of the possible new information which could be introduced about the topic should be selected: thus it also plays a part in determining how the topic should be treated, but in terms of content rather than of rhetoric.

'Perspective' is the only element in this system of analysis which will be somewhat unfamiliar. 'Perspective' is both a semantic and a pragmatic concept: semantically perspective refers to the model structure by which the individual defines her world and her place within it; therefore it determines the possible ways that the individual can act, think and speak. Pragmatically perspective determines the appropriateness of the discourse and is defined in terms of context, i.e., the point of view, attitudes and so on of all discourse participants. The implication of this is that what is asserted must be seen as appropriate to the intentions and goals of the writer. However, in a discontinuous discourse such as a writing test, the interpretation of the speech act may not be the same for initiator and respondent, or for the respondent and the participant who follows up. The interpretation will depend on the perspective of each participant. But in an academic or specific academic writing test question, the perspective of the initiator, the test constructor, has great power because of the social context, and this perspective is imposed upon the test taker. It is one of the 'rules of the game', and the test taker is expected to know that and to respond appropriately. But as we saw in the previous section not all testees do respond appropriately, and the reason is not always pragmatic incompetence or some other incompetence.

Perspective is closely related to 'outcome space', a concept which is central to Pollitt et al's model, and which is explained in detail in

Pollitt and Hutchinson (n.d.). When the constructor of a writing test designs a question, he has clear ideas about the kinds of responses he expects and those he is prepared to accept as legitimate: this is the test constructor's outcome space. The design of the question, and the delineation of the outcome space is inevitably influenced by the constructor's perspective, since perspective influences everything we do. But since testee writers also bring a perspective to the writing test question, there is a problem that the writer's outcome space may not perfectly match that of the test constructor. When the writer perceives, or believes she perceives, what the test constructor's perspective is, but finds what is asserted to be untrue or inappropriate within her own perspective, conflict results between the awareness of the conventions of the writing test as a discourse within a fixed power structure, and the need of the writer to state that which she believes to be true within her world view. The claim here is that it is this conflict which generates challenges.

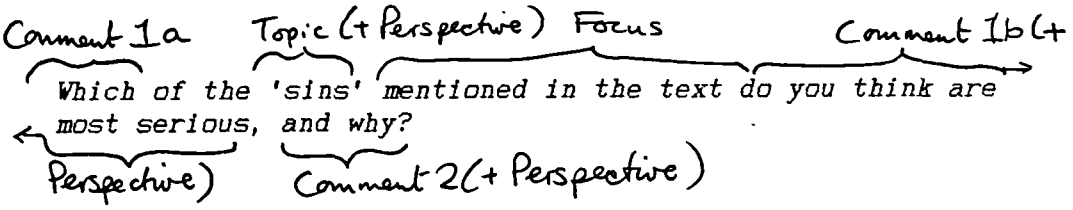
As we saw earlier, the problem is exacerbated when the writer's response to the test question is not followed up by the initiator but by a third participant. This participant, the rater, will also bring a perspective to the activity, and it may not perfectly match that of either the test constructor or the writer, although given the life experience and social role of the rater his perspective is likely to be more similar to that of the test constructor than the writer. The rater, then, brings to the evaluation of answers his own ideas about the kinds of responses which are acceptable, that is, he has his own outcome space. We saw in Chapter 5 that some raters found particular answers or material within answers "irrelevant" while others did not: this is a difference in outcome space.

3.2. Application of the task analysis categories to tasks on the three writing tests

The identification of these categories of task variables is context-bound and tentative: it is not possible at this stage to quantify them. Presence or absence of a particular variable in a task does not make that task either good or bad, easy or difficult. It was suggested in section 2, for example, that the resources for SAPQ SS were too long and that some writers had problems processing them to choose appropiate material for an answer, while the resources for M2Q1 PS provided almost nothing in the way of useable material for an answer to the question. Both ends of this continuum seem to be difficult: the 'ideal' is presumably somewhere in between.

3.2.1. GAPQ

The GAPQ question, marked up to show its structure, is as follows:



The resources for the task (see Appendix A3) were perhaps a little difficult linguistically, and were misinterpreted by a number of students. The possibility of a problem with the rubric was referred to in sub-section 2.1.3. We also saw earlier that the lexical item 'sin' in the question caused some problems.

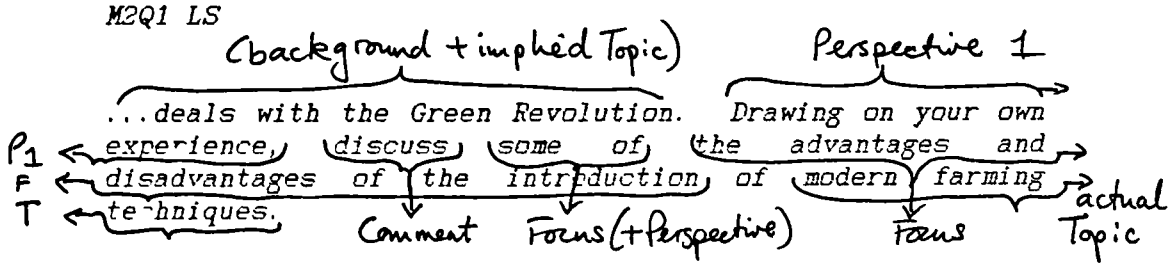
There was little help with an answer structure in the resources, and although they provided input to the topic of an answer, they were not very helpful with focus: on the other hand, there was little possibility of plagiarism - lexical items could reasonably be transferred but not larger chunks. The question provides a partial structure for an answer

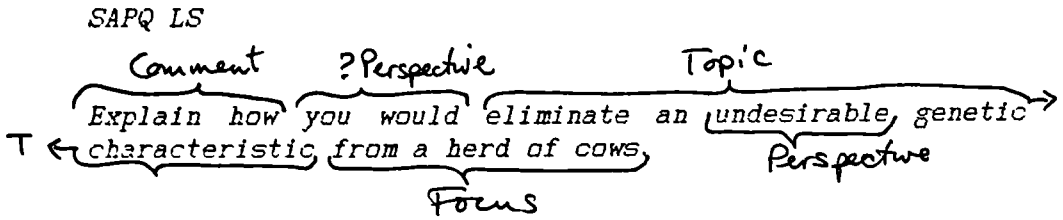
through the two parts of the comment ("which...why?"), suggesting an organisation based on a paragraph per 'sin' and moving from worst to least serious: it is also linguistically fairly easy, except for the problem with 'sin' already noted. On the other hand, there is little help with content for an answer in the question.

Although the outcome space is quite wide there is a perspective which assumes that belief in sin as a concept is reasonable, and some writers challenged this perspective. Additionally, although the subject of good and evil was certainly familiar to all the writers, the linking of these concepts with 'sin' is not a universal, and 'sin' as a subject may be more difficult and unfamiliar. The subject matter was mainly rather abstract, but the inclusion of various numerical data gave a false impression of concreteness. The required mode of discourse was argument based on personal experience, and this seemed to be a mode which caused little difficulty.

3.2.2. The LS questions

The LS questions, marked up to show their structures, are as follows:





3.2.2.1. M2Q1 LS

Despite the combination of a quotation from Swift and an extract from a textbook, the resources did not seem to be misinterpreted, but they were often plagiarised. It was in this Module that the greatest problem of students not doing this question justice by comparison to the second arose: the second was on the carbon cycle, evidently a very familiar topic for most of these writers, and a number of them wrote unduly long and detailed answers to that question at the expense of this one. The question was often misinterpreted as being about the green revolution, due to the lead-in reference which seems to suggest that the green revolution and modern farming methods are synonymous.

The subject matter seemed familiar to most writers, as far as it is possible to judge from the written responses; it was also concrete, and expressed concisely in semi-technical rather than highly technical language. The close relationship between the resources, the question and the answer made the question easy to answer but encouraged plagiarism. This was exacerbated by the strong structural parallel between resources and answer and by the help provided in determining rhetorical structure by the question.

The mode of discourse was primarily expository, since writers were asked to discuss advantages and disadvantages but were not asked to state, or state and defend, a position. Writers were also asked to draw on their own experience, i.e., the intention was that answers would be constructed from a personal perspective, but this did not occur.

The question appears to express an assumption that a writer will have a perspective which enables her to find support in her experience for an implicit argument that modern farming methods have advantages and disadvantages in every context. This may not, however, be the writer's perspective: some writers may see only advantages because of a strong

economy and advantageous agricultural conditions in their own country, while others may see only disadvantages because their own country has a weak economy and poor agricultural conditions which are better suited to alternative agricultural technology. Thus the perspective implied by the question narrowed the outcome space in a way which writers may not have been able to accept: equally, they may not have had a breadth of experience which would enable them to address both sides of the issue from personal experience. Writers did not overtly challenge, however: they appeared to ignore the perspective requirement altogether. Some raters were stricter than others in taking seriously the expectation that writers should bring in their own experiences.

3.2.2.2. SAPQ LS

The use of visuals coupled with short pieces of text seemed to ensure that the resources were not misinterpreted. There were no observed problems with the rubric or the question. There was no evidence of unfamilarity with the subject matter of this task and some evidence of extreme familiarity by some writers. The subject matter and the resources, with clear layout, many diagrams and short, separate sections of text were concrete and straightforward. The mode of discourse was expository ("explain how") with no additional mode requirements, and suggested a fairly simple, linear, text structure for an answer. This was the structure used in the resources, which simplified the construction of an answer but also proved a great temptation for plagiarism.

There was some inappropriate plagiarism because the question transformed the resources slightly, requesting the method for eliminating an undesirable characteristic, whereas some less skilled writers followed the resources closely in describing genetic inheritance more generally.

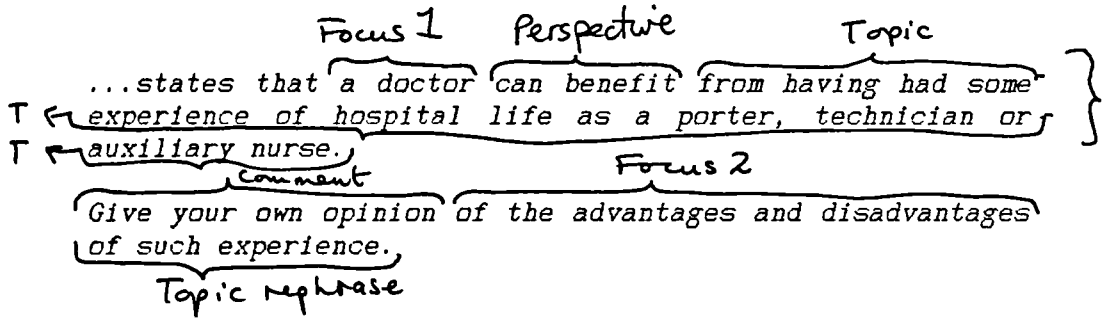
Linguistically the question did not appear to pose problems despite the lexical items 'eliminate' and 'undesirable': these seemed to be familiar to

the writers. Although the outcome space suggested by the structure of the question is quite proscribed, there were no challenges or incompetencies; writers seemed to share the initiator's perspective and see this as a valid task.

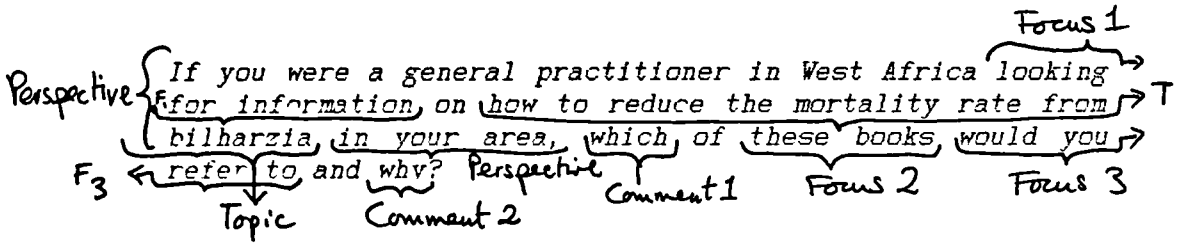
3.2.3. ME questions

The ME questions, marked up to show their structure, are as follows:

M2Q1 ME



SAPQ ME



3.2.3.1. M2Q1 ME

There was some misinterpretation of the resources for this task, some writers apparently being unable to distinguish main and supporting information, but it did not result in seriously inappropriate answers. There were no rubric problems: the second question on this Module was simple and required a short response. There were some misinterpretations of the question, which was treated by some writers as expository rather than argumentative. The subject matter was carried in a text which had

an unhelpful structure and which moved from abstract to concrete, general to specific in poorly marked ways.

Although the context was very familiar to these doctors, the arguments were probably less so, and they picked the concrete parts of the text and paid little attention to the more abstract parts. The rhetorical mode expected of an answer apparently is a mixture of exposition ("advantages and disadvantages") and personal argument ("give your own opinion"), and this was handled awkwardly in most cases. Writers tended to include all the points made in the resources and briefly mention their own position on each of them, rather than constructing a cohesive argument. This may also have been due to the weak relationship between the structure of the resources and the best structure for an answer, so that the right content was hard to find.

The question was not particularly difficult linguistically, but the phrasing of the question with its apparent emphasis on the balance of pro and con suggested a narrower outcome space than was in fact the case.

3.2.3.2. SAPQ ME

There were no observed problems with the resources, rubric or question of this task. The resources consisted solely of a medical bibliography, and there was no other content which could be used in an answer. The writers all appeared to be familiar and at ease with reading medical references, and made appropriate choices. We may see this subject matter as concrete and easy but limited in extent, requiring additional input from the writer.

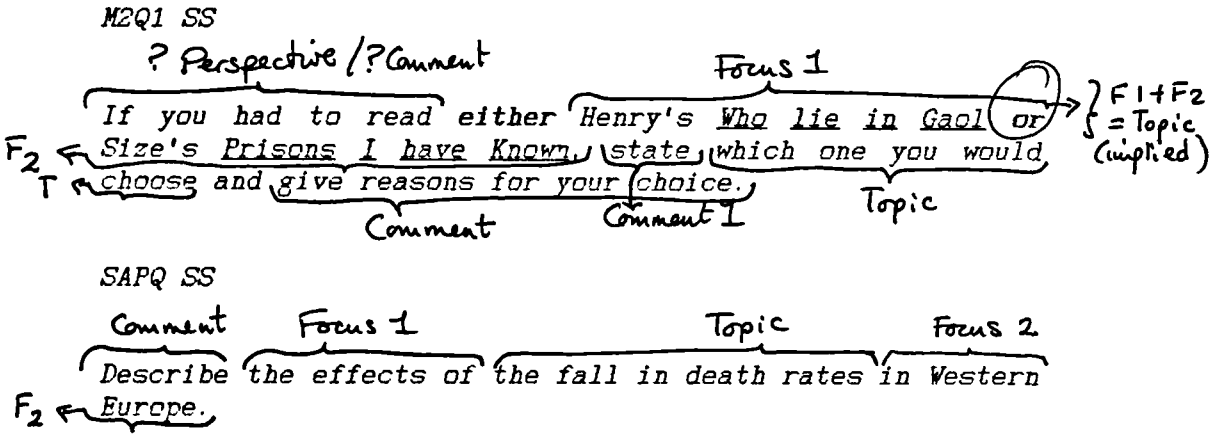
Unlike other questions, this one offers a context and purpose for the answer. The mode of discourse required appears on first consideration to be, like GAPQ, a personal argument, but is in fact a reasoned exposition,

and exposition from the writer's professional role as a doctor. While the outcome space may seem very constrained, there were no challenges or evidences of incompetence: doctors appeared to view this as a valid task.

The question is linguistically quite difficult and is rhetorically complex, using the hypothetical conditional, several focuses, and a strong specification of perspective: it did not, however, appear to cause comprehension problems for the writers.

3.2.4. SS questions

The SS questions, marked up to show their structure, are as follows:



3.2.4.1. M2Q1 SS

The only misinterpretation of the resources apparent here was the belief of a number of writers that Henry was the author's first name. It caused no serious problems, since writers who made that mistake seemed to be so culturally uninformed that they continued to think of Henry as a woman. There were no observed misinterpretations of the rubric or question.

It was striking how many writers declared, or otherwise showed themselves, to be completely unfamiliar with the subject matter of prisons, and especially of women in prisons. The resources proved difficult because they were so limited, especially in relation to the focus: there was little support in the resources to help the writer build an argument in favour of one book over the other. Although the resources, such as there are, are fairly concrete, the task is somewhat abstract: this seems to be the reverse of the usual situation in these tasks.

We might describe the required mode of discourse as personal argument, except that there is nothing in the task to help the writer build the kind of personal investment in a decision of one book or the other which is necessary for a convincing personal argument answer. Few answers were convincing. If there is little support for the content of an answer, there is no support for the structure of one; the little support offered by the introductory paragraph to the bibliography was so urgently needed by writers that many plagiarised it wholesale.

Linguistically the question is a relatively simple one, but the outcome space it offers is so limited, and so apparently unreasonable, requiring a perspective where no perspective or material to manufacture one exists, that writers often did poorly and many challenged.

3.2.4.2. SAPQ SS

There were signs of misinterpretation of these resources: the text was long and fairly difficult, and its subject was wider than that of the question, so careful reading and selection was needed. In contrast, the rubric and question appeared to generate no problems.

Many writers seemed rather unfamiliar with the subject matter beyond a vague idea of what the term 'death rates ' meant (and some appeared not

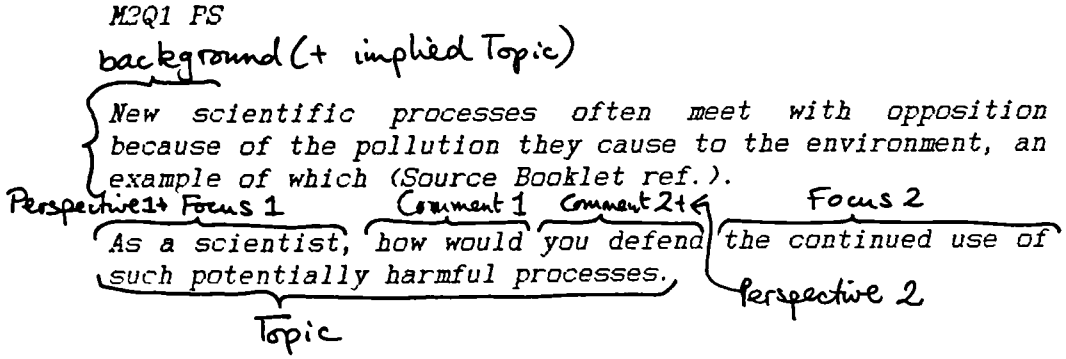
to have even that much knowledge). The required mode of discourse, exposition, was not complicated by any additional mode expectations, and seemed to be handled adequately. There was a strong relationship between the question and the resources, but the appropriate content was well embedded into the much longer text and for some writers the search was very difficult. Once appropriate content was located there appeared to be a temptation for some writers to reward themselves by plagiarising. There was only a weak relationship between the resources and the question structure, and it may be because of this, coupled with the difficulty of information retrieval, that some answers were disorganised.

The question was not difficult linguistically, but the outcome space was quite narrow. Although there were no challenges, writers apparently considering the task to be valid, if difficult, there is a sense of discomfort in answers which suggests that this <u>is</u> a hard task.

3.2.5. PS questions

The PS questions, marked up to show their structure, are as follows:

M2Q1 PS



SAPQ PS

Discuss some of the ways in which air pollution can be reduced.

3.2.5.1. M2Q1 PS

In this small corpus (7 writers) one writer wrote about the composition of the air, one wrote a short answer to this question and a longer one to the second question, and the others made more or less overt challenges.

In the less overt challenges it would be possible to assume that the writer had misinterpreted the question and failed to notice the instruction to "defend", but the rate of occurrence of challenges was so high that seems unlikely. The resources were very short, and were not on the topic or focus of this question: the "example" mentioned in the question is really a passing mention of sulphur compounds from factories, which are described as "impurities". There is very little support for an answer in the resources in terms of either content or structure, although this appeared to be a familiar subject to most writers. The required mode of discourse is argumentative, but not really a personal argument. The writer is asked to write from within her role "as a scientist".

There are, then two perspectives here which seriously restrict the outcome space: the writer must defend rather than oppose or 'discuss' these processes, and she must do so on scientific grounds. For many writers this appeared to cause a real role conflict between the role of participant in an academic test and all they know about the 'rules' of that game, and the role of scientist, i.e., member of a specific disciplinary, professional community with its own mores. We have looked in some detail at the challenges which resulted in section 1.

3.2.5.2. SAPQ PS

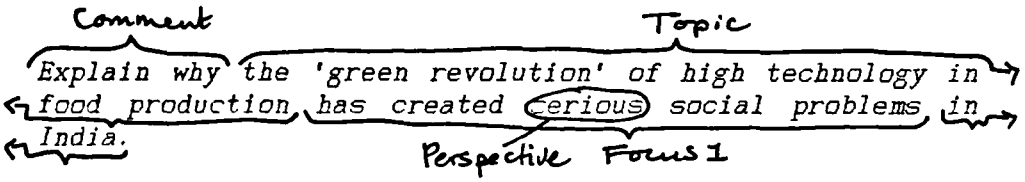
There were no observed problems of resources, rubric or question with this task. The resources were the same as for M2Q1 PS, and the text is fairly concrete and not difficult. The required mode of discourse was exposition, with no additional mode expectations. Writers needed to

generate additional content to develop that in the resources but the resources provided a good basis. There was not much help there with structure for an answer, however.

The question is linguistically easy and was in general well handled; the outcome space is quite wide, and the design of the task appears to encourage writers to go beyond the resources to their specialist knowledge.

3.2.6. GA questions

The M2Q1 question is the same as the M2Q1 SS question and is not repeated here: the SAPQ GA question, marked up for structure, is as follows:



There was one misinterpretation of the resources, which was discussed in sub-section 2.1.2.; there were no observed misinterpretations of the rubric or question. Many writers gave the impression of being familiar with the subject matter, but one or two appeared to have no prior concept of the 'green revolution' at all. The resources seeemed to provide most writers with sufficient background to feel familiar with both the topic and the focus, and provided some useful content. Plagiarism was a temptation, however, due to the relative ease of the text, its concreteness, and the close relationship between certain parts of it and the question. Help with structuring an answer was also available from the resources.

The question was linguistically easy, and although the outcome space was narrowed by the perspective, which assumed that writers would agree the

'green revolution' had led more to problems than to benefits in India, that was the perspective also taken by the text and writers seemed to find it valid. There were no overt challenges.

3.3. Task equivalence and predicting scores

It can be seen from the foregoing discussion that the tasks in these three writing tests are not equivalent in design. First, they are not equivalent across Modules: that is, the five M2Q1 questions discussed are not equivalent to each other. They do not all require the same mode of discourse; they do not all require the same amount of personal investment; they do not all provide the same amount of content from which to build an answer; they do not all offer the same amount of support with structure for an answer; some generate many more challenges than others. The same is true, although it would appear on the task analysis above, to a lesser degree, of the SAPQ questions across Modules. We might predict from this that questions will not be equivalent in difficulty.

We may predict from the discussion above that the questions will yield different score levels across Modules. Comparisons of scores across Modules are not possible, however, because there is no claim nor any evidence that the writers in each Module were equivalent in any sense: this was not an experimental study. Indeed, on the evidence of the ELTS overall scores (ELTSOA) presented in Chapter 4 it seems clear that they are not equivalent. We may, however, compare scores within Modules and relate these to the task analysis above.

The discussion of the LS tasks suggests that M2Q1 LS will be more difficult than SAPQ LS. The mean scores support this: M2Q1 mean is 5.585 while SAPQ LS mean is 6.293. The LS group received the lowest mean score on ELTSOA apart from the GA group, but their mean scores on the three writing tests place them second in rank order of writing performance for

GAPQ and M2Q1, and first for SAPQ. If we were to assume a 'flat' profile of skills (an assumption we cannot in fact make) this would suggest that both LS writing tasks were relatively easy. In fact the LS mean score for M2Q1 is very close to their ELTSOA mean, suggesting that perhaps the explanation is that writing tasks in other Modules are relatively difficult.

A similar pattern can be predicted for M2Q1 ME and SAPQ ME, although some difficulty was identified with both these tasks. The mean scores are close together: M2Q1 ME is 5.909 and SAPQ ME is 6.000, not supporting a prediction that the tasks will both be shown to be rather difficult. The ME group's mean ELTSOA score placed them first in rank order of the Modular groups, as did their GAPQ and M2Q1 writing test mean scores. Their SAPQ mean score was the same as their ELTSOA, but LS outperformed them on SAPQ. Again, if we were to assume a 'flat' profile, it would appear that LS SAPQ at least is relatively easy. However, we saw in Chapter 4 that there was an inverse and weak correlation between the scores on the SAPQ/M2Q1 tasks, suggesting that one or several task variables are generating quite different kinds of responses in these doctors, one task favouring some and the other task favouring others. Such a pattern of correlations suggests that for individuals flat profiles would not be found. The most striking differences of task variables between the two questions were: text length and type; mode of discourse; degree of role specification; specification/non-specification of purpose for task.

The discussion of the SS tasks leads to a prediction that M2Q1 SS will yield lower scores than SAPQ SS. This is not in fact the case: the M2Q1 SS mean is 5.555 and the SAPQ SS mean is 4.926. The length of the text for SAPQ SS and the embeddedness of the relevant content within it led, as was noted, to short answers. Since there are strong indications that length is often a significant variable in explaining scores in holistic readings this may be the explanation. Clearly the weakness of the M2Q1

SS task was outweighed by the resource difficulty and the failure to provide an answer structure in the SAPQ SS task. The SS group exhibited generally lower scores on the writing tests than their ELTSOA mean score would have predicted on the assumption of a 'flat' profile. On ELTSOA the SS group was ranked second: on M2Q1 they were ranked third, and on GAPQ and SAPQ they were ranked lowest. In none of the writing tests did the group's mean score come close to their mean score on ELTSOA, suggesting either that all the writing tasks are relatively difficult, or that SS students tend to have 'marked' profiles with relatively weak writing skills.

We saw in the discussion of the PS tasks that M2Q1 PS generated many challenges, many of them strong and lively. We also saw that raters responded to these challenges in different ways: different from rater to rater and different from writer to writer. Although we might predict that such a constraining task will cause problems for writers and lead to low scores, examination of the answers and of the raters' responses suggested that some challenges were made with confidence and conviction, and that some at least were valued by raters. It becomes more difficult on closer examination to make a prediction. In fact M2Q1 PS has a mean score of 5.428 while SAPQ PS has a mean score of 5.143: apparently writers were not as disadvantaged by the narrow outcome space as we might have predicted. Perhaps the controversiality of the question, coupled with the fact that it is a question physicists often have to deal with, perhaps unfairly in their view, engages them in the topic powerfully (as SAPQ PS clearly did not), and it is this engagement which raises their M2Q1 scores. It does not, however, affect their writing performance relative to other Modular groups: the PS group were ranked third on ELTSOA and on GAPQ: they were ranked fourth on both M2Q1 and SAPQ.

In the GA Module we predicted that the problematic M2Q1 task would result in lower scores for that task than for the SAPQ task, and were shown to be wrong. The same prediction is made here for the use of that

task in the GA Module, but it is made with more conviction because the SAPQ GA task appeared to have no serious problems with task variables, unlike the SAPQ SS task. In fact the mean score for SAPQ GA is 5.826 and the mean score for M2Q1 GA is 5.261, confirming the prediction. The GA group exhibited the lowest mean ELTSOA score, and were also placed lowest on M2Q1, with a mean very close to their ELTSOA: they were ranked fourth on GAPQ and third on SAPQ. Considering the close correlations between writing test scores for this group it would seem that SAPQ GA in particular is an easy question.

At this level then, the system for task analysis appears to allow prediction with some success but also some error. Clearly, there are other factors operating which have not been identified in the task analysis.

3.4. What makes a task SAP rather than GAP?

In the task analysis above no particular attention was paid to identifying SAP rather than GAP task variables. Reviewing the discussion of incompetence in section 2 it would be difficult to find any variable there which could be identified as existing only in SAP contexts: problems of misinterpretation occur on all writing tests, and problems of pragmatic incompetence occur on all academic writing tests. In the discussion of challenges we saw that challenges occur when writers cannot value the task, and it was suggested earlier in this section that challenges are more frequent when outcome space is limited. Challenges to GAPQ seemed to be on what we might describe as a 'moral' basis: that is, the writers did not seem to orient their personal/cultural morality in the same underlying values as the question implied. M2Q1 PS generated some similar challenges but also challenges which seemed to be on a 'disciplinary' basis: we saw that writer number 122 wrote from the perspective of a scientist and gave a scientists's reasons for challenging the values implied by the question. We also saw how writer

number 93 challenged the M2Q1 SS task from the perspective of her role within her discipline. Through their challenges, PS writers were able to transform the M2Q1 PS task into one they could value: this was less easy to do with GAPQ.

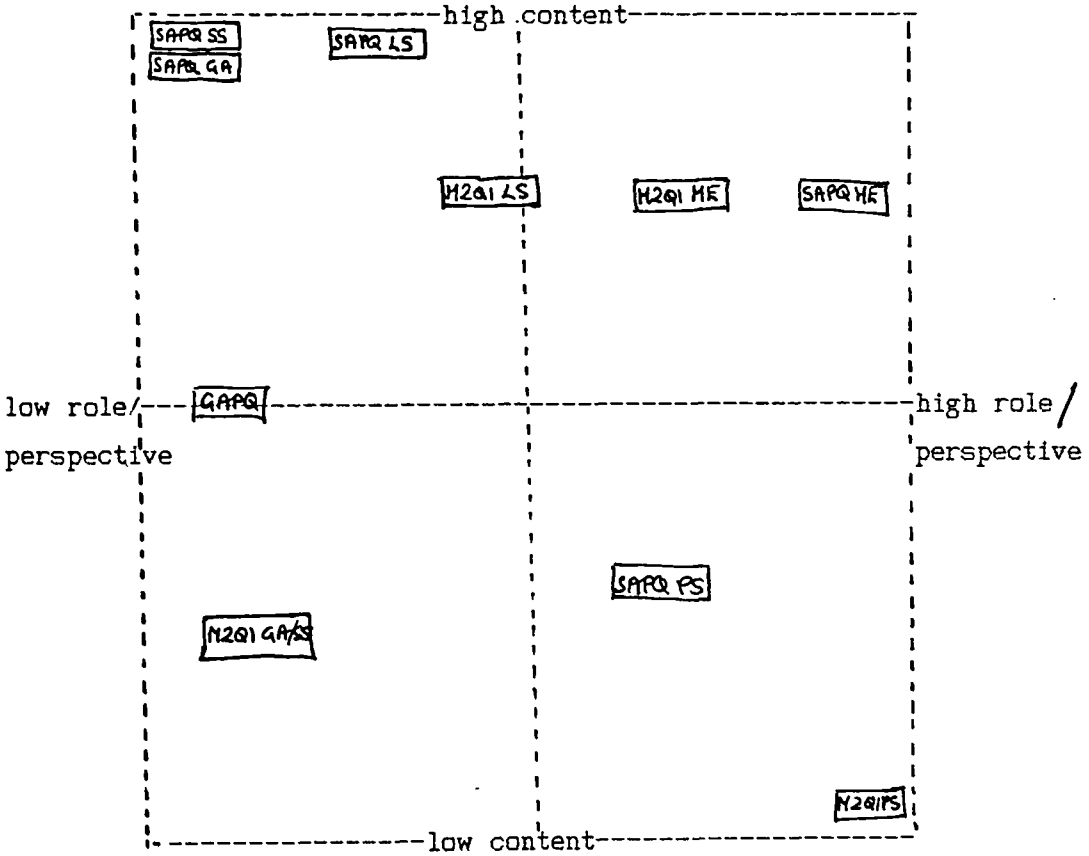
We would expect a central SAP variable for a writing test task to be the need to respond from within a discipline: that is, if role and perspective are assigned they should be authentic ones for a member of that disciplinary community. M2Q1 PS and SAPQ ME are particularly strong in that regard. M2Q1 ME assumes a fairly narrow set of roles from within which the writer may write, but does not limit the perspective, although many writers interpreted it as doing so. The M2Q1 LS task attempts to offer the writer the chance to write from within her own perspective by asking for personal experience, but as we saw earlier this was ignored by writers: the explanation suggested is that by asking writers to discuss both advantages and disadvantages the personal role and perspective is taken away from them. The perspective offered for the M2Q1 SS/GA task was one of compulsion, and role was neither stated nor could be inferred: on this variable this is not a SAP task.

Another central SAP variable is content. A SAP writing test can be expected to be based firmly within the writer's content knowledge and require her to demonstrate that knowledge. Further, according to studies by Houghton (1984), James (1984), Johns (1985) and Swales (1982), and as confirmed by this researcher's faculty survey reported earlier, faculty within disciplines in their writing tests design tasks which emphasise the selection of relevant data from sources and the reorganisation of those data to build an appropriately organised response. All the SAPQ tasks were designed to attempt to fit this design variable, but in the cases of ME and PS this was not easy to achieve within the constraints of the source material. In each of these cases the task requires the writer to work with the resources but to go beyond them, into her knowledge of her discipline. Thus they move from content to role. Of

the M2Q1 questions only LS and ME are much related to content; little content is available for SS/GA and virtually none for PS. As we have seen, writers on M2Q1 PS are able to move from content to role, even when this involves a challenge, but on M2Q1 SS/GA the task is too far from most writers' perceptions of their role, either personal or disciplinary, for this to be possible.

Figure 6.3.1. is an attempt to show how these two variables interrelate for each M2Q1 and SAPQ question to make each of them a more or less SAP task:

Figure 6.3.1.: Dimensions of SAPness/GAPness for M2Q1/SAPQ



3.5. What makes a writer's response SAP rather than GAP?

It appears from the detailed study of answers and tasks in the corpus used for this study that the first and foremost aspect of a writer's response which characterises it as SAP rather than GAP is that the writer approaches the task from a disciplinary perspective, writing from within a disciplinary role. Here we are approaching the concept of 'voice' as this is used in creative writing. The answers to SAPQ tasks by writers 18 (LS), 31 (PS) and 35 (ME) in Appendix I are each examples of answers which are written with the 'voice' of the discipline, that is, of a writer writing with the authority of their field. The answers by writers 1 (GA) and 121 (PS) are included in Appendix I as examples which do not have the voice of disciplinary authority.

Our investigations of task variables have made clear that certain elements of the question either facilitate or impede the emergence of a 'voice' on any particular task. The topic and focus of the question must fit the writer's perception of a legitimate arena for discourse in the discipline (even if in fact it isn't) so that it may be valued. The comment instruction(s) must be reasonable within the discourse of the discipline: it would seem that all interactional modes are acceptable to writers in this corpus, but that a comment requiring a personal (i.e., outside the role of member of a discipline) response makes writers uneasy; if the writer sees that the personal response is required to a public issue (as in the case of M2Q1 PS) this is accepted but answers tend to be from the role of society member rather than disciplinary community member. The combination of interactional and personal writing called for in M2Q1 LS did not seem to enable writers to find an authoritative voice. The perspective of the question plays a major part in determining the outcome space of an answer, and therefore of the voices with which it will be legitimate to use in responding. The reading of answers suggests that questions with an implicit rather than explicit perspective are most successful, sending a message to writers

who are firmly grounded in their disciplines that it is appropriate to respond from the disciplinary perspective, while not preventing writers who are at the point of entry into the discipline from finding a more general stance from which to respond. Both ME questions seem to be particularly successful in this regard, but of course this is easiest with ME since the community of testees is narrowest and most clearly defined in the minds of all participants in the discourse - test constructor, writer/medic and rater.

The writer's possession of sufficient basic knowledge of the content area of the task of the writing test is also central to making her response SAP rather than GAP. The most convincing answers appeared to take the content of the answer and develop an answer from it not simply by selection, reorganisation and paraphrase, but by addition of content (factual content rather than opinions). It can be seen in Appendix I that this was convincingly done by writer 14 (M2Q1 LS) and writer 110 (SAPQ LS); writer 120, a writer with weak linguistic skills, also showed convincing use of her own content knowledge on SAPQ PS. Writers without that personal familiarity with the content produced answers which worked within the information provided.

The impression resulting from close study of the responses in this corpus is that an incompetent answer can never be a SAP answer, but a challenge answer may be - and that in fact challenges are most likely to occur when a low SAP task combines with a writer who is firmly grounded in the role, perspective and content of her own SAP. The cases of writers 31 and 122 on M2Q1 PS, discussed in sub-section 2.2.3., are interesting in this regard. Writer number 31, in challenging, appears to reject her role as a scientist, and responds from the perspective of a member of the wider community; writer number 122 obeys the directive to respond as a scientist, but rejects the perspective of either defending or opposing potentially harmful processes as outside her role as a scientist. We saw that raters did not greatly value writer 122's answer.

We appear to have been able to find some elements of SAPness in some of the 'SAP' tasks, and some instances of SAP responses from writers. There also seem to be 'SAP' tasks which are much closer to being GAP tasks on Figure 6.3.1.: M2Q1 SS/GA appears to be more centrally a GAP task on these dimensions than GAPQ is. There also are responses from writers which have little to mark them as SAP responses, either of 'voice' or of content addition and transformation.

We saw in the final section of Chapter 5 that the raters gave few signs of taking a SAP perspective themselves while rating, and thus we cannot assume that there will be any match between SAPness in a response and SAPness in a score; further, we found that the scoring procedure had no specifically SAP-focussed criteria. It is not likely that writers who are potentially SAP writers will produce SAP responses to tasks that are only uncertainly SAP at best. But if they do, there is little to indicate that such responses can be handled in valid ways through the scoring criteria and procedure, or that raters are intuitively operating their SAP criteria and valuing SAP responses.

CHAPTER SEVEN

CONCLUSIONS

1. General academic purpose writing tests or specific academic purpose writing tests?

The study reported here began with the attempt to investigate the claim that writing tests in a student's academic discipline ('SAP' writing tests) will yield more meaningful information about the student's writing proficiency in English than writing tests applicable to the university community at large ('GAP' writing tests). Although limited support for a strong ESP construct was found, the study has not provided any strong argument for the use of SAP rather than GAP writing tests in the assessment of the writing of postgraduate students seeking entry to British tertiary education.

Although when the subjects were treated as a single group both SAP mean scores were significantly higher than the GAP mean scores, and correlations among the scores were all significant, only 38% at best of the score variance could be accounted for. In only two of the five Modular groups studied, LS and SS, were mean scores for SAP writing tests significantly higher than mean scores for the GAP writing test. Across all groups, correlations showed that performances of individuals varied from question to question, and that the variation was not explained by a consistent SAP/GAP distinction. The only group for which the two SAP writing tests met the criterion for parallel forms was GA. GA was the only case in which scores on the two SAP questions correlated more highly with each other than scores on either SAP question correlated with GAPQ, and the only case in which correlations among all three tests were equally significant.

Evidently, SAP questions are lifting scores, but not always for the same students. This pattern could be explained theoretically by a hypothesis in which students with strong skills in the particular subject area would be more advantaged by any specific academic question in that subject area than students with weaker skills in the same area, but it does not provide support for a simplistic assertion that SAP writing tests will necessarily advantage students correctly placed on a six way division of disciplines. Students may be correctly placed, but not have the necessary skills of the discipline, or not have mastered the actual content necessary for the question asked.

Further, in two cases, PS and SS, the two SAP questions were not significantly correlated with each other, while each of them was significantly correlated with GAPQ. For the subjects treated as a single group, and for the other three Modular groups, the correlation between GAPQ and one of the SAP writing tests was higher than the correlation between the two SAP writing tests. In the PS Module, SAPQ and GAPQ met the criterion for parallel forms, which is contrary to all predictions. A finding of closer relationships between one of the SAP questions and the GAP question appears to confound the hypothesis of a SAP advantage and requires an alternative explanation. The pattern noted suggests that in many cases M2Q1 is more a 'GAP' task than a 'SAP' task. Few of the correlations were able to account for as much as 60% of the score variance, suggesting that an explanation might be found in aspects of test design rather than in the test subjects. There was no case in the study where the data fit a pattern which could be predicted by an ESP construct.

At the completion of the empirical study it was only possible to say that no predictions had been fulfilled except the prediction of significant relationships among the three writing tests, and even this prediction was not fulfilled in every case. The two SAP questions on the SS Module were not significantly related, and none of the questions on the ME Module

were significantly related, despite the fact that this Module showed the most similar mean scores across questions. In this Module, individuals were performing differently on the three questions but the resulting totals were very similar. Either ME students are more sensitive to small changes in tasks than other students, or the differences among the tasks for this group were more dramatic than the differences among tasks for other students.

It was concluded that other factors were having strong influences on the data and making it impossible to approach a valid measurement of the variables in which we are interested. For this reason the further studies were carried out into the four principal task variables on a writing test: the task, the scoring procedure, the raters, and the writers. The purpose of these further studies was to understand more about the causes of variation in the scores, and to search for ways of reducing, or at least of predicting and accounting for, such variation.

2. Validity of the scoring procedure

The scoring procedure used for scoring all three writing tests for each group was the original M2Q1 scoring procedure. There was no evidence to suggest that raters found the procedure more appropriate to the scoring of some questions than of others. This procedure was found to be unreliable when used operationally, and development was carried out which made it more reliable under operational conditions, and which permitted the concurrent investigation of the procedure's validity. No evidence was found to suggest that the procedure became more valid: raters gave no indication that they were applying SAP criteria when using the procedure for its stated purpose, i.e. for scoring answers to supposedly SAP questions. Not only was content not a criterion, it was specifically excluded and discounted as a criterion. None of the criteria elicited from raters' rating behaviour displayed any SAP characteristics, and as

the scoring procedure was made explicit and refined neither raters nor administrative bodies involved commented on this lack.

The motive force for the development of the scoring procedure was the need for improved reliability: no attempt was made to move the procedure towards a SAP basis. The development work on the scoring procedure showed that the criteria used for the scoring of M2Q1, a supposedly SAP question, demonstrated no features which could be viewed as specific to SAP writing and inapplicable to GAP writing. If anything, the situation is the reverse, i.e., the scoring procedure for M2Q1 is in fact a GAP procedure, being applied to a purportedly SAP writing test.

3. Validity of raters' rating processes

The study of the raters in action when rating M2Q1 suggested that it is possible to use observations of what raters do to articulate criteria more exactly and to aid raters in identifying what it is they are responding to in order that they may do so more consistently. It confirmed that raters were not applying SAP criteria in making their judgements, or rewarding SAP responses when they recognised them in testees' writing. Rather, in the instances where it seemed that a SAP basis for a judgement might be appropriate it was seen that raters either failed to demonstrate awareness of the possibility of such a basis, or veered away from basing their judgement in such criteria. Thus, disagreements between raters as to the appropriateness of responses occurred most often in the areas of content accuracy and relevance.

A certain disjunction was observed between raters' apparent ability to follow the instruction to ignore content and their insistence on relevance. Raters seemed to be concerned with organisation and ideas within a single genre and not to have varying expectations from question to question across Modules; they referred often to 'message', but this seemed more to do with the fulfilment by the writer of certain text

structural expectations than with the absolute worth of the content chosen: relevance was as much a matter of form as of fact. The advisability of basing questions on the specific material of provided texts was also called into question by the frequent occurrence of plagiarism or other misuse of the input texts. It was clear that the rating processes of these raters were based in GAP expectations of writing in academic settings.

4. What writers' responses revealed about tasks

Study of the answers produced by writers enabled the identification of a number of response variables which appeared closely related to task variables and through which tasks could be examined. Content gave clear indications of being a SAP task variable, but content requirements alone did not appear to be sufficient to stimulate writers to produce responses which deserved to be characterised as SAP rather than paraphrases/summaries of the content of a text. A content requirement is not necessarily a call for the writer to refer to the related text: in fact answers which remained close to the input material seemed unconvincing and ran the risk of plagiarism. SAPQ LS was an exmaple of this. Tasks which made content demands, but which assumed that the writer could herself provide some content, seemed to result in convincing answers, for example, M2Q1 ME.

Similarly, role/perspective gave clear indications of being a SAP variable, but those questions which limited these requirements to a demand that the writer bring in personal experience were not particularly convincing to the writers as tasks, for example M2Q1 LS. Only those tasks in which the writer was able to make a personal investment succeeded in producing a response in a recognisable SAP voice, and in this regard M2Q1 PS was particularly successful, contrary to prediction. The identification of this interaction of content and role/perspective to create a unique outcome space for each writer brings the task close to

'personal' writing. We are accustomed to thinking of academic writing, and particularly specific academic writing, as impersonal, as 'expository'. From this study it would appear that the most convincing SAP writing is in important ways also 'personal' writing.

When tasks were analysed to discover which of them appeared to be based more heavily in those dimensions which yielded SAP responses, it appeared that the two ME SAP questions combined the most features of SAP content and role, with SAPQ ME making greater role demands. SAPQ LS required high content engagement but made rather low role demands, while M2Q1 LS was more centrally SAP with moderate role demands and fairly high content requirements. No other questions were unequivocally SAP: SAPQ PS was borderline, with low content and fairly high role involvement. M2Q1 PS was on one outer periphery, with very high role demands and no content, while SAPQ SS and GA were on the other, with very high content expectations and no role. The other questions fell clearly into the GAP half of the diagram.

5. What has been learned?

The studies carried out make it clear that if a writing test is to be a valid test of specific academic writing, a scoring procedure must be designed which makes explicit this intention and which states, describes and illustrates the valid criteria to be applied to the judgement of SAP responses to SAP tasks. Nothing was found to suggest that valid SAP scores can result from the application of a GAP scoring procedure. Whether or not experienced teachers of English as a second/ foreign language are able to conduct valid SAP ratings, given proper training and guidance, was not an issue explored in this study, although Hamp-Lyons (1986) suggests that they are. Certainly they do not have, ready-made in their minds, SAP criteria which they can apply uninstructed, nor can they intuitively recognise and reward SAP responses.

There was little to suggest that raters were consistently recognising and valuing authentic SAP responses from writers. Indeed, on occasion raters appeared to view SAP responses with suspicion. The disjunction between the implicit claims of the test and the expectations of the writers, and the basis of judgements in the scoring procedure and the raters, may explain why score levels and correlation patterns could not be well predicted. It seems clear that all the aspects of the studies of the four key variables of task, writer, scoring procedure and rater are interacting in complex ways, and that these interactions were poorly understood at the time of test construction.

It also seems fairly clear from these studies that the constructors of ELTS M2 had no principled basis for the construction of truly SAP writing test tasks. No consistent design parameters can be reconstructed from the M2Q1 tasks apart from the appeal to personal experience, and this is a parameter which is patently not particular to SAP tasks. Further, there is no evidence that the constructors had any design principles or parameters in mind at all, apart from the appeal to experience just mentioned and the need to link the writing task to an input text chosen for another purpose, i.e., a text used previously to test reading comprehension. Some tasks show some SAP characteristics but this is haphazard and unpredictable, appearing to be more due to luck, or good intuitions, than to judgement.

When tasks are at best hesitantly SAP, raters are GAP and the scoring procedure is GAP, it would be foolish to expect meaningful distinctions to occur between responses to those tasks labelled SAP and those labelled GAP. There is little evidence here to indicate that if a question did generate a SAP response it would be given a SAP reading and scoring. The evidence of these studies suggests that, while there may be some advantage to some testees in a SAP question rather than a GAP question, differences in difficulty level of questions have more influence in determining scores. While the study reported in Chapter 6 did not

succeed in identifying all sources of task difficulty it did show some progress, enabling, for example, GAPQ and SAPQ SS to be identified as difficult, and SAPQ LS as easy. Study of challenges led to some incorrect predictions, but also led to a greater understanding of the way in which the writer's sense of role interacts with the content possibilities of a task to give to the writer's response a sense of engagement with the task. It led also to the observation that challenges based on the writer's role as a member of a disciplinary community were found less acceptable by these raters than challenges based on the writer's role as a member of a wider community. This observation returns us to the interaction between all the key variables in a writing test, and the unlikelihood of establishing consistent relationships between tasks on the basis of their degree of 'SAPness' as long as every key variable is not functioning in ways which are valid for the test purpose.

It has not been possible to establish a clear advantage for SAP writing test tasks over GAP writing test tasks, nor has it been possible to establish the validity of the parameters of specific disciplines as these have been used in the ELTS. The failure of SAP questions purportedly in the same disciplinary area to correlate highly may be wholly the result of poor task design, which has been clearly established, or it may be partly the result of invalidity of the discipline-specific construct upon which the test is based. Since specifications of the precise basis upon which the boundaries between Modules in ELTS were drawn are not available, tasks cannot be compared to any *a priori* construct. What does seem to have been established, however, is that without exact specifications of the parameters of specific disciplines ('Modules') and of the construct of writing as it occurs in that discipline, it will not be possible to show conclusively that supposedly SAP questions in any discipline are more appropriate and can yield more accurate scores than supposedly GAP questions.

6. SAP writing tests, GAP writing tests, and the fulfilment of expectations

We have seen that it is possible for M2Q1 to fulfil reliability expectations, but that to date there are no indications of the fulfilment of the validity expectations of a specific academic writing test. Although the test may claim a certain face validity, tasks are of dubious and erratic content and construct validity; neither the scoring procedure nor the raters' processes are valid. There must be doubt whether the responses made by writers are valid SAP responses, given the failure of validity of the tasks. There is no consistent set of results which can be explained through a construct of specific academic writing. While the backwash from the test may be beneficial, there have been no studies to establish this, and the concerns about test security which have been heard relate to the negative backwash created if test items are leaked and 'stock answers' become common. The same criticisms can be levelled at SAPQ. It is clear that SAP writing tests are less efficient than GAP writing tests, and on this ground, without any strong factors to indicate otherwise, it appears that no argument can be made for specific academic writing tests over general academic writing tests.

This is not, however, to claim that general tests of writing in academic settings are necessarily better than specific ones. All writing tests have validity problems. It is difficult to design writing tests of any kind which can reflect the construct of writing as we characterised it in Chapter 1.

We know that writing, composing, is a process which occurs over time, but we give little time for writing tests. We know that few good composed products are recognisable in their first draft form, but writing tests require and judge what are inevitably first drafts. We know that writing is a process of discovery and of learning, but we require writing tests in order to measure what it is that writers already know. We know that

writing is purposive and interactional, but we construct writing tests which assume that the testing purpose and the interaction with the tester will be sufficient stimulus for the writer to treat the activity of responding as a communicative act. The only feature of the model of composing which was presented in Chapter 1 which occurs in these writing tests appears to be the normative feature, and this detracts from rather than adds to validity as long as we are in the position of not possessing a sound basis of understanding of the norms of written discourse in each discipline against which writers' test essays can be judged.

There is nothing to indicate that at present tests of writing in academic settings, general or specific, approach construct validity. However, as we understand more about what it means to write in academic settings, and how this writing differs from discipline to discipline, it will become more feasible to design writing tests with construct valdity in those terms.

The failure of ELTS M2 to meet the expectations set up for it in Chapter 4 does not stand alone. The same criticisms apply equally to SAPQ and to GAPQ. Writing tests designed and implemented as these were can confidently be predicted to fail to meet any stringent expectations. The development of any writing test must begin with a detailed *a priori* construct validation which looks closely at the setting which the test is intended to operate in and to reflect. A scoring procedure must be developed to match the construct. Tasks must also be developed to test the construct and must be carefully pre-tested to ensure that they are on target. Raters must be carefully trained to understand not only the scoring procedure *qua* procedure, but also the construct underlying it. Scores resulting from the test must meet not merely reliability expectations but more importantly validity expectations.

7. Future research and development

It must be clear from the foregoing that more work is needed in all these areas. A good deal of research is already being done into ESP and the characterisation of specific academic communities and their discourse, and all this research will inform test development. There is no evidence as yet, for example, that the way in which the ELTS divides up the disciplinary 'cake' has any construct validity: the limited research which has been done suggests that it does not.

Much work is also being done into the development of scoring procedures for writing tests, and this can be applied to the problems of scoring writing tests for the kinds of general and specific academic settings with which the ELTS is concerned. After a period in which holistic scoring reigned uncontested in writing assessment, a shift toward a range of carefully designed, valid analytic schemes, of which the multiple trait procedure developed for the third version of the M2 assessment is but one, can be generally discerned.

Many researchers are concerned with the apparent disjunction between process and product in writing, and attention is increasingly being paid to the ways in which that disjunction can be lessened, so that product demands take better account of processes and processes are shaped toward an awareness of the products which are their eventual goal. Some of this research is centred in academic communities, general or specific.

These research areas need to be linked together so that our understanding of how we can assess writing performance, process and product, in academic settings, will be increased. There are as yet few signs that this linking research is under way, and it will take a good many replicatory studies in a range of academic contexts, building as time goes by on the developments made in each of the areas separately, to enable us to reach that understanding.

There are two aspects of the research reported here which do not appear to be receiving attention at the present time, which have considerable potential for increasing our understanding of the writing assessment process in any context, including academic contexts, and which can be pursued independently of other research developments.

First, the ethnographic study of raters in the process of rating student answers provided a rich source of data. This source of data has been neglected in almost all research in the testing of writing, or else data collection methods such as self-reports have not permitted a clear understanding of what it is that raters actually do while judging writing. There is much to be learned from such studies. The ethnographic study here was limited, because the data were collected during the development of a scoring procedure: further studies in less constrained contexts would be fruitful. It would also be useful to replicate such studies, using the same answers and different raters, in order to investigate the effect of the development of a rating community on the score patterns which evolve.

Second, the reconstruction of task variables from writers' responses provides another rich source of data. This study was limited by the small number of tasks, but analysis of many tasks in this way will teach us a great deal about what it is in tasks which writers respond to and which makes them more or less difficult for different groups of writers. Ideally this would be coupled with structured interviews of the writers immediately after they have completed the writing task and the interviewer has read the answer, in order to collect self reports about responses, interpretations and processes. A study of writers from the same discipline functioning as a community, reading and discussing their own and others' answers soon after taking the test would also provide interesting information: writers asked to respond to each others' writing would reveal their own criteria for judgements in ways which would parallel the ethnographic study of raters.

Finally, while this study has not provided a vindication of those who would test the writing proficiency of applicants for tertiary education places through specific academic rather than general academic writing tests, it has not shown that such an aim is mistaken in principle. It has shown only that such an aim is impractical in the present state of our knowledge. The study has provided a foundation from which research and development of tests of writing in a variety of academic settings, general and specific, may go forward.

BIBLIOGRAPHY

Achiba, M. & Y. Kuromiya. 1983. 'Rhetorical patterns extant in the English compositions of Japanese students'. JALT Journal 5 : 1-13.

Alatis, J, ed. 1969. Linguistics and the Teaching of Standard English to Speakers of Other Languages or Dialects. Monograph Series on Language and Linguistics No. 22. Georgetown University Press: Washington D.C.

Alderson, J.C. 1981a. 'Report of the discussion on testing English for Specific Purposes' in Alderson & Hughes, eds.

Alderson, J.C. 1981b. 'Report of the discussion on General Language Proficiency' in Alderson & Hughes, eds.

Alderson, J.C. & A. Hughes, eds. 1981. Issues in Language Testing: ELT Documents No. 111. The British Council: London.

Alderson, J.C. & A.H. Urquhart. 1983. 'The effect of student background discipline on comprehension: a pilot study' in Hughes & Porter, eds.

Alderson, J.C. & A.H. Urquhart. 1984. 'Do we need ESP tests?' Paper presented at TESOL Convention, Houston, TX.

Allen, H.B. 1965. Teaching English as a Second Language. McGraw-Hill: New York.

Allen, J.P.B. & S.P. Corder. 1974. Techniques in Applied Linguistics. Vol.3 in Edinburgh Course in Applied Linguistics. Oxford University Press: Oxford.

Allen, J.P.B. & A. Davies. 1977. Testing and Experimental Studies. Vol.4 in Edinburgh Course in Applied Linguistics. Oxford University Press: Oxford.

American Psychological Association. 1966. Standards for Educational and Psychological Tests. American Psychological Association: Washington, D.C.

Anastasi, A. 1976. Psychological Testing. Macmillan: London.

Anderson, C.C. 1960. 'The new STEP essay test as a measure of composition ability'. Educational and Psychological Measurement 20: 95-102.

Anderson, P., C. Muller & J. Brockmann, eds. 1984. New Essays in

Technical and Scientific Communication: Theory, research and criticism. Vol. 2 in Baywood Series in Technical and Scientific Communication. Baywood: Farmingdale, NY.

Anderson, R.C., R.J. Spiro & M.C. Anderson. 1978. 'Schemata as scaffolding for the representation of information in connected discourse'. American Education Research Journal 15: 433-440.

Andrich, D. 1973. Latent trait psychometric theory in the measurement and evaluation of essay writing ability. PhD. University of Chicago.

Angelis, P. 1983. 'Performance validation in tests of second language proficiency'. Paper presented at TESOL Convention, Toronto, ON.

Applebee, A.N. 1983. 'Writing and learning in school settings' in Nystrand,ed.

Applebee, A.N. & G. Brossell. 1985. Research on writing assessment'. NTNW Notes, Nov 1985.

Arapoff, N. 1970. 'Writing: a thinking process'. English Teaching Forum 3: 4-8.

ATESOL. 1973. Testing in Second Language Learning: New Dimensions. ATESOL: Dublin.

Bachman, L.F. & A.S. Palmer. 1980. 'Basic concerns in test validation' in Alderson & Hughes, eds.

Bachman, L.F. & A.S. Palmer. 1981. 'Some comments on the terminology of language testing'. Paper presented at Language Proficiency Assessment Symposium, Rosslyn, VA.

Bachman, L.F. & A.S. Palmer. 1982. 'The content validation of some tests of communicative proficiency'. TESOL Quarterly 16: 449-465.

Bailey, F.G. 1977. Morality and Expediency. Blackwell: Oxford.

Bailey, R.W. 1983. 'Literacy in English: an international perspective' in Bailey & Fosheim, eds.

Bailey, R.W., R.T. Brengle & E.L. Smith, Jr. 1980. 'Measuring student writing ability' in Freedman & Pringle, eds.

Bailey, R.W. & R.W. Fosheim, eds. 1983. Literacy for Life. Modern Language Association: New York.

Baker, E.L. & J.L. Henman. 1983. 'Task structure design: beyond linkage'. Journal of Educational Measurement 20: 149-164.

Ballard, B. & J. Clanchy. 1986. Literacy in the university: an 'anthropological' approach. Mimeo.

Ballard, P.B. 1923. The New Examiner. University of London Press: London.

Banman, R. & J. Sherzer, eds. 1974. Explorations in the Ethnography of Speaking. Cambridge University Press: London.

Barnard, P.J., P. Wright & P. Wilcox. 1979. 'Effects of response instructions and question style on the ease of completing forms'. Journal of Occupational Psychology 52: 209-226.

Baron, D. 1975. 'Non-standard English, composition, and the academic establishment'. College English 37: 176-183.

Barritt, L.S. & B.M. Kroll. 1978. 'Some implications of cognitive-developmental psychology for research in composing' in Cooper & Odell, eds.

Bar-Lev, Z. 1986. 'Discourse processes and "contrastive rhetoric"'. Discourse Processes 9: 235-246.

Basich, R. & L. Polin. 1985. 'Research designs for studying writing assessment'. NINW Notes, Nov. 1985.

Beach, R. & L.L. Bridwell. 1984. New Directions in Composition Research. Guilford Press: New York.

Basso, K. 1974. 'The ethnography of writing' in Banman & Scherzer, eds.

Bazerman, C. 1981. 'What written knowledge does: three examples of academic discourse'. Philosophy of Social Sciences 11: 361-382.

Beaugrande, R. de. 1979. 'Moving from product toward process'. College Composition and Communication 30: 357-363.

Beaugrande, R. de. 1983a. 'Psychology and composition: past, present and future' in Nystrand, ed.

Beaugrande, R. de.. 1983b. 'Linguistic and cognitive processes in developmental writing'. International Review of Applied Linguistics 21: 125-144.

Beaugrande, R. de. 1984a. Text production: Toward a science of

composition. Vol. XI in Advances in Discourse Processes Series. Ablex: Norwood, N.J.

Beaugrande, R. de. 1984b. 'Forward to the basics: getting down to grammar'. College Composition and Communication 35: 358-367.

Becher, A. 1981. 'Towards a definition of disciplinary cultures'. Studies in Higher Education. 6: 109-122.

Becher, A. Forthcoming. 'The disciplinary shaping of the profession' in Clark, ed.

Ben-Amos, D. & K.S. Goldstein, eds. 1975. Folklore = Performance and Communication. Mouton: The Hague.

Bennett, N. 1973. Research Design. Open University Press: London.

Bereiter, C. & M. Scardamalia. 1983. 'Levels of inquiry in writing research' in Mosenthal et al, eds.

Bernstein, R.S. & B.R. Tanner. 1977. The California High School Proficiency Examination: evaluating the writing samples. ERIC ED 147806.

Biggs, J.B. & K.F. Collis. 1982a. Evaluating the Quality of Learning: The SOLO Taxonomy. Academic Press: London.

Biggs, J.B. & K.F. Collis. 1982b. 'The psychological structure of creative writing'. American Journal of Education 26: 59-70.

Biglan, A. 1973. 'The characteristics of subject matter in different academic areas'. Journal of Applied Psychology 57: 195-203.

Bishop, A. 1978. The concern for writing. Educational Testing Service: Princeton, N.J.

Bizzell, P. 1982. 'College composition: Initiation into the academic discourse community'. (Essay Review) College English 12: 191-207.

Black, J.B., D. Wilkes-Gibbs & R.W. Gibbs, Jr. 1983. 'What writers know that they don't know they need to know' in Nystrand, ed.

Bloom, B.J 1972. Taxonomy of Educational Objectives: the classification of educational goals. Vol. 1. Cognitive Domain. Longman: London.

Bloom, B.J. 1973. Taxonomy of Educational Objectives: the classification of educational goals. Vol. 2. Affective Domain. Longman: London.

Bloom, B.S. 1971. Handbook on Formative and Summative Evaluation of

Student Learning. McGraw-Hill: New York.

Bloomfield, L. 1933. Language. Henry Holt: New York.

Bobrow, D.G. & A. Collins, eds. 1975. Representation and understanding: Studies in cognitive science. Academic Press: New York.

Bracewell, R.J. 1983. 'Investigating the control of writing skills' in Mosenthal et al, eds.

Braddock, R., R. Lloyd-Jones & L. Schoer. 1963. Research in Written Composition. National Council of Teachers of English: Urbana, IL.

Bradley, H. 1919. On the Relationship Between Spoken and Written Language with Special Reference to English. Oxford University Press: Oxford.

Brand, A.G. & J L. Powell. 1985. 'Emotions and the writing process: a description of apprentice writers'. Journal of Educational Research 79: 280-285.

Brannon, L. & C.H. Knoblaugh. 1982. 'On students' rights to their own texts: a model of teacher response'. College Composition and Communication 33: 157-166.

Branthwaite, A., M. Trueman & T. Berrisford. 1981. 'Unreliability of marking: further evidence and a possible explanation'. Educational Review 33: 41-46.

Breland, H.M. & J.L. Gaynor. 1979. 'A comparison of direct and indirect assessments of writing skill'. Journal of Educational Measurement 16: 119-128.

Breland, H M. & R.J. Jones. 1982. Perceptions of Writing Skill. ETS Research Report No. 82-47. Educational Testing Service: Princeton, N.J.

Bridgeman, B. & S.B. Carlson. 1983. Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students. TOEFL Research Reports No. 15. Educational Testing Service: Princeton, N.J.

Bridgeman, B. & S.B. Carlson. 1984. 'Survey of Academic Writing Tasks'. Written Communication 1: 247-280.

Briere, E. 1975. 'Current Trends in Second Language Testing' in Palmer & Spolsky, eds.

Briere, E. & F. Hinofotis, eds. 1979. Concepts in Language Testing. TESOL: Washington, D.C.

British Council, The. 1978. English for Specific Purposes. ELT Documents 101. English Teaching Information Centre, The British Council: London.

British Council, The. 1980. Team Teaching in ESP. ELT Documents 106. English Teaching Information Centre, The British Council: London.

British Council, The. 1981. The ESP Teacher: role, development and prospects. ELT Documents 112. British Council English Language and Literature Division: London.

British Council, The/University of Cambridge Local Examinations Syndicate. 1982. The English Language Testing Service: First Report. English Language Testing Service, British Council: London.

British Council, The/University of Cambridge Local Examinations Syndicate. 1982. Administrator's Manual. English Language Testing Service, British Council: London.

British Council, The/University of Cambridge Local Examinations Syndicate. 1983. The English Language Testing Service: Second Report. English Language Testing Service, British Council: London.

British Council, The/University of Cambridge Local Examinations Syndicate. 1985. Assessment Guide for M2 Writing. English Language Testing Service, British Council: London.

British Council, The/University of Cambridge Local Examinations Syndicate. Forthcoming. New Assessment Guide for M2 Writing. English Language Testing Service, British Council: London.

British Council, The/University of Cambridge Local Examinations Syndicate. no date. User Handbook. English Language Testing Service, British Council: London.

British Council, The/University of Cambridge Local Examinations Syndicate. no date. English Language Testing Service: An Introduction. English Language Testing Service, British Council: London.

Britton,J.N. 1950. 'The meaning and marking of imaginative compositions'. New Era 31: 7-12.

Britton, J.N. 1963. 'Experimental marking of English compositions written by fifteen year olds'. Educational Review (Birmingham) 16.

Britton, J.N. 1978. 'The composing processes and the functions of writing' in Cooper & Odell, eds.

Britton, J.N., T. Burgess, N.C. Martin, A. McLeod & H. Rosen. 1975. The Development of Writing Abilities (11-18). Macmillan: London.

Britton, J.N., N.C. Martin & H. Rosen. 1966. Multiple Marking of English Compositions: an account of an experiment. Schools Council Exams Bulletin, 12. H.M.S.O.: London.

Brookes, A. & P. Grundy. 1985. 'Activating the learners' contribution in the development of academic writing skills'. Paper presented at the SELMOUS Conference, Reading, England.

Brooks, V. 1984. CSE English Examinations: an evaluation of the procedures employed by the East Midland Regional Examinations Board to assess oral and written English. PhD, University of Leicester.

Brooks, V. 1980. Improving the Reliability of Essay Marking. A Survey of the Literature with Particular Reference to the English Language Composition. MEd, University of Leicester.

Brossell, G. 1986. 'Current research and unanswered questions in writing assessment' in Greenberg et al, eds.

Brossell, G. 1983. 'Rhetorical specification in essay examination topics'. College English 45: 165-174.

Brossell, G. & B. Hoetker Ash. 1984. 'An experiment with the wording of essay examination topics'. College Composition and Communication 35: 423-425.

Brown, H.D. 1976. Language Learning Special Issue No.4: Papers in Second Language Acquisition. Research Club in Second Language Acquisition: Ann Arbor, Michigan.

Brown, R. 1981. 'The need for better information' in Fredericksen & Dominic, eds.

Brown, S. 1980. What Do They Know? A Review of Criterion-Referenced Assessment. H.M.S.O.: London.

Bruce, B., A. Collins, A. Rubin & D. Gentner. 1978. A cognitive approach

to writing. Technical Report No. 89. Center for the Study of Reading: Urbana, Illinois.

Burgess, T. 1980. Outcomes of Education. Macmillan: London.

Burgess, T.C. & N.A.F. Greis. 1970. English language proficiency and academic achievement among students of English as a second language at the college level. ERIC ED 074 812.

Burhans, C.S. Jr. 1983. 'The teaching of writing and the knowledge gap'. College English 45: 639-656.

Burke, K. 1945. A Grammar of Motives. Prentice-Hall: Englewood Cliffs, N.J.

Burt, C. 1940. The Factors of the Mind. London University Press: London.

Burton, N.W. 1978. 'Societal standards'. Journal of Educational Measurement 15: 263-271.

Butler, M., A. Casmier and others. 1984. Special Issue: Students' Right to their Own Language. College Composition and Communication 25.

Butterfield, C.J. 1945. The Effect of a Knowledge of Certain Grammatical Elements on the Acquisition and Retention of Related Punctuation Skills. PhD, University of Iowa.

Buxton, E.W. 1958. An Experiment to Test the Effects of Writing Frequency and Guided Practice upon Students' Skill in Written Expression. PhD, Stanford University.

Campbell, D.T. & D.W. Fiske. 1959. 'Convergent and discriminant validation by the multi-trait - multi-method matrix'. Psychological Bulletin 56: 81-105.

Canale, M. 1982. 'Evaluating the coherence of student writing in L1 and L2'. Paper presented at the Colloquium on Discourse Analysis and Language Learning, TESOL Convention, Honolulu, Hawaii.

Canale, M., N. Frenette & M. Belanger. 1983. 'On the interdependence of L1 and L2 in student writing in a minority setting'. Paper presented at the Annual Language Testing Research Colloquium, Ottawa, ON.

Canale, M. & M. Swain. 1979. Communicative Approaches to Second

Language Teaching and Testing. Ontario Ministry of Education: Ontario.

Canale, M. & M. Swain. 1980. 'Theoretical bases of communicative approaches to second language teaching and testing'. Applied Linguistics 1: 1-47.

Candlin, C.N. 1977. 'Communicative language teaching and the debt to pragmatics' in Rameh, ed.

Candlin, C.N., J.M. Kirkwood & H.M. Moore. 1978. 'Study skills in English: theoretical issues and practical problems' in Mackay & Montford, eds.

Carlman, N. 1984. The effects of scoring method, topic, and mode on grade 12 students' writing scores. Dissertation Abstracts International 42 02A.

Carlson, S. & B. Bridgeman. 1986. 'Testing ESL student writers' in Greenberg et al, eds.

Carlson, S.B., B. Bridgeman, R. Camp & J. Waanders. 1985. Relationship of Admission Test Scores to Writing Performance of Native & Nonnative Speakers of English. ETS Research Reports 19. Educational Testing Service: Princeton, N.J.

Carmines, E.G. & R.A. Zeller. 1979. Reliability and Validity Assessment. Sage Publications: Beverly Hills, CA.

Carpenter C. & J. Hunter. 1981. 'Functional exercises: improving overall coherence in ESL writing' TESOL Quarterly 15: 425-434.

Carre, C. 1984. 'Using language actively in science classrooms' in James, ed.

Carroll, B.J. 1980. Testing Communicative Competence: An Interim Study. Pergamon Press: Oxford.

Carroll, B.J. 1981. 'Specifications for an English Language Testing Service' in Alderson & Hughes, eds.

Carroll, J.B. 1941. 'A factor analysis of verbal abilities'. Psychometrika: 6: 279-308.

Carroll, J.B. 1961. Testing the English Proficiency of Foreign Students. Center for Applied Linguistics: Washington, D.C.

Carroll, J.B. 1965. 'Fundamental considerations in testing the English language proficiency of foreign students' in Allen, ed.

Carroll, J.B. 1967. 'Foreign language proficiency levels attained by language majors near graduation from college'. Foreign Language Annals 1: 131-151.

Carroll, J.B. 1973. 'Foreign language testing: will the persistent problems persist?' in ATESOL.

Carroll, B.J. & S.M. Sapon. 1959a. Modern Language Aptitude Test. A. Psychological Corporation: New York.

Carroll, B.J. & S. Sapon. 1959b. Modern Language Aptitude Test Manual. Psychological Corporation: New York.

Cartier, F. 1975. 'Criterion-referenced testing of language skills' in Palmer & Spolsky, eds.

Carver, D. 1983. 'Some propositions about ESP'. ESP Journal 2: 131-137.

Cast, B.M.D. 1939-40. 'The efficiency of different methods of marking English compositions'. British Journal of Educational Psychology 9: 257-269 & 10: 49-60.

Cazden, C.B., V.P. John & D. Hymes. 1972. Functions of Language in the Classroom. Teachers College Press: New York.

Chafe, W L. 1976. 'Givenness, contrastiveness, definitiveness, subjects, topics and point of view' in Li & Thompson eds.

Chafe, W.L. 1979. 'The flow of thought & the flow of language' in Givon, ed.

Chafe, W.L., ed. 1980. The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production. Vol. III in Advances in Discourse Processes Series. Ablex: Norwood, N.J.

Chai, S.H. & P.L. Woehlke. 1979. 'The predictive ability of standardized tests of English as a foreign language' in Silverstein, ed.

Chamberlain, R.G.D. & M.K.S. Flanaghan. 1978. 'Developing a flexible ESP programme design' in British Council.

Chaplen, E.F. 1970. The Identification of Non-Native Speakers of English Likely to Underachieve in University Courses Through Inadequate Command of the Language.PhD, University of Manchester.

Chaplen, E.F. 1971. 'The reliability of the essay sub-test in a university entrance test in English for non-native speakers of English' in Perren & Trim, eds.

Charney, D. 1984. 'The validity of using holistic scoring to evaluate writing'. Research in the Teaching of English 18: 65-81.

Chase, C.I. 1968. 'The impact of some obvious variables on essay test scores'. Journal of Educational Measurement 5: 315-318.

Chater, P. 1984. Marking and Assessment in English. Methuen: London.

Chaudron, C. 1983. 'Evaluating writing - effects of feedback on revisions'. Paper presented at TESOL Convention. Toronto; Ontario.

Chickering, A.W. ed. 1981. The Modern American College. Jossey-Bass: San Francisco.

Christenson, F. 1963. 'A generative rhetoric of the sentence'.College Composition and Communication 14: 155-161.

Christenson, F. 1965. 'A generative rhetoric of the paragraph'. College Composition and Communication 16: 144-56.

Christenson, F. 1967. Notes Toward a New Rhetoric: Six Essays for Teachers. Harper & Row: New York.

Christopher, B. 1985. 'Holistic scoring and reader thinking'. Paper presented at NTNW Notes, Nov 1985.

Cipolla, C.M. 1969. Literacy and Development in the West. Penguin: Harmondsworth

Clancy, P.M. 1980. 'Referential choice in English & Japanese narrative discourse' in Chafe, ed.

Clark, B.R. ed. Forthcoming. The Academic Profession. University of California Press: San Francisco.

Christenson, F. 1967. Notes Toward a New Rhetoric: Six Essays for Teachers. Harper & Row: New York.

Clark, J.L.D. 1978. 'Psychometric considerations in language testing' in Spolsky, ed.

Clark, M. 1980. 'There is no such thing as good writing (so what are we looking for ?)' in Freedman & Pringle (1980a), eds.

Clark, S.O. 1980. 'The correlation between aptitude scores and achievement measures in Japanese and German' in Oller & Perkins, eds.

Claunch, N.C. 1964. Cognitive & motivational characteristics associated with concrete & abstract levels of conceptual complexity. Joint ONR/NIMH Report. Princeton University: New Jersey.

Clyne, M. 1981. 'Culture & discourse structure'. Journal of Pragmatics 5: 61-66.

Coffman, W.E. 1971a. 'On the reliability of ratings of essay examinations in English'. Research in the Teaching of English. 5: 24-36.

Coffman, W E. 1971b. 'Essay examinations' in Thorndike, ed.

Cohen, A.M. 1973. 'Assessing college students' ability to write compositions'. Research in the Teaching of English. 7: 356-371.

Cohen, A.M. 1980. Testing Language Ability in the Classroom. Newbury House: Rowley, Mass.

Cohen, A.M. & F.B. Brawer. 1983. 'Functional literacy for community college students' in Bailey & Fosheim, eds.

Cole, P. & J.L. Morgan, eds. 1975. Syntax and Semantics: Vol. III,Speech Acts. Academic Press: New York.

College Entrance Examinations Board. 1946. Forty Sixth Annual Report of the Executive Secretary. College Entrance Examination Board: New York.

Condon, E.C. 1975. 'The cultural context of language testing' in Palmer & Spolsky, eds.

Conlan, G. 1978. How the Essay in the College Board English Composition Test Is Scored. English Testing Services: Princeton, N.J.

Connor, U. 1984. 'Argumentative patterns in student essays: cross-cultural differences'. Paper presented at AILA Conference, Brussels.

Cook, L. 1946. 'Teaching grammar and usage in relation to speech and writing'. Elementary English Review 23: 193-198.

Cook-Gumperz, J. & J.J. Gumperz. 1980. 'From oral to written culture: the

transition to literacy' in Whiteman, ed.

Cooper, C. 1985. Aspects of Article Introductions in IEEE Publications. MSc, University of Aston, Birmingham.

Cooper, C.R. 1977. 'Holistic evaluation of writing' in Cooper & Odell eds.

Cooper, C.R., ed. 1981a. The Nature and Measurement of Competency in English. National Council of Teachers of English: Urbana, IL.

Cooper, C. 1981b. 'Competency testing: issues and overview' in Cooper (1981a).

Cooper, C.R. 1983. 'Procedures for describing written texts' in Mosenthal et al, eds.

Cooper, C.R. 1985. 'New approaches to primary trait scoring'. NTNW Notes Nov 1985.

Cooper, C.R. & L. Odell, eds. 1977. Evaluating Writing: Describing, Measuring, Judging. National Council of Teachers of English: Urbana, Illinois.

Cooper, C.R. & L. Odell, eds. 1978. Research on Composing: Points of Departure. National Council of Teachers of English: Urbana,Illinois.

Cooper, M.M. 1984. 'The pragmatics of form: how do writers discover what to do when ?' in Beach & Bridwell, eds.

Cooper, P.L. 1984. The Assessment of Writing Ability: A Review of Research. ETS Research Report 84-12. Educational Testing Service: Princeton, N.J.

Cooper, R.L. 1968. 'An elaborated language testing model' in Upshur & Fata, eds.

Corbett, E.P.J. 1967. 'What is being reviewed?'. College Composition and Communication. 18· 166-172.

Corder, S.P. 1973. Introducing Applied Linguistics. Oxford University Press: Oxford.

Coulthard, M., ed. 1986. Talking About Text. English Language Research: Birmingham, England.

Coulthard, M. & M.C. Ashby. 1975. 'Talking with the doctor'. Journal of Communication 25: 240-247.

Cowie, A.P. & J.B. Heaton. 1977. English for Academic Purposes. BAAL/SELMOUS: London.

Cowan, G. & E. Cowan. 1980. Writing. John Wiley & Sons: New York.

Cox, R. 1968. Examinations and Higher Education: A Survey of the Literature. Society for Research into Higher Education Ltd.

Crawford, A.B. & P.S. Burnham. 1946. Forecasting College Achievement. Yale University Press: Yale.

Cresswell, M.J. & J.G. Houston. 1983. Norm and criterion referencing of performance levels in tests of educational attainment. Mimeo. Associated Examining Board.

Criper, C. 1981. 'Reaction to the Carroll paper (2)' in Alderson & Hughes, eds.

Criper, C. & A. Davies. 1983. ELTS Validation Project Outline (revised April 1983). Mimeo. University of Edinburgh.

Criper, C. & A. Davies. 1986. Final Report of the Edinburgh ELTS Validation Project. Mimeo. University of Edinburgh.

Cronbach, L.J. 1961. Essentials of Psychological Testing. Harper & Row: New York.

Cronbach, L.J. 1971. 'Test validation' in Thorndike, ed.

Cronbach, L.J. 1976. 'Equity in selection - where psychometrics and political philosophy meet'. Journal of Educational Measurement 13: 31-41.

Cronbach, L.J. & P.E. Meehl. 1972. 'Construct validity in psychological tests' in Noll et al, eds.

Crowhurst, M. & G.L. Piche. 1979. 'Audience and mode of discourse effects on syntactic complexity in writing at two grade levels'. Research in the Teaching of English 13: 101-109.

Culhane, T., C. Klein-Braley & D. Stevenson. 1982. Practice and Problems in Language Testing IV Department of Language and Linguistics Occasional Papers 26, University of Essex.

D'Angelo, F. 1974. 'A generative rhetoric of the essay'. College Composition and Communication25: 388-396.

Daly, J. & M. Miller. 1975. 'The empirical development of an instrument

to measure writing apprehension'. Research in the Teaching of English 9: 242-249.

Darnell, D.K. 1968. The Development of an English Language Proficiency Test of Foreign Students Using a Cloze-Entropy Procedure. ERIC ED 024 039

Davey, A. 1974. The Formalization of Discourse Production. PhD, University of Edinburgh.

Davies, A. 1964. English Proficiency Test Battery, Version A. British Council: London.

Davies, A. 1965a. Proficiency in English as a Second Language. PhD, University of Birmingham.

Davies, A. 1965b. 'Language proficiency testing' in Report on Sixth Meeting of International Conference on Second Language Problems. ETIC, The British Council: London.

Davies, A. 1967. 'The English proficiency of overseas students'. British Journal of Educational Psychology 37: 165-174.

Davies, A. 1968. Language Testing Symposium. Oxford University Press: Oxford.

Davies, A. 1973. 'Language proficiency testing and the syllabus' in ATESOL.

Davies, A. 1977. 'The construction of language tests' in Allen & Davies, eds.

Davies, A. 1978 (July & October). 'Language Testing Survey Articles' in Language Teaching and Linguistics Abstracts.

Davies, A. 1981. 'Reaction to the Bachman & Palmer and the Vollmer papers' in Alderson & Hughes, eds.

Davies,A. 1983a. An Evaluation of the ELBA Test. (Report to the University of Edinburgh Senatus). Mimeo. University of Edinburgh.

Davies, A. 1983b. 'The validity of concurrent validation' in Hughes & Porter, eds. (1983b).

Denman, M.E. 1978. 'The measure of success in writing'. College Composition and Communication 29: 42-46.

Diederich, P.B. 1967. Cooperative Preparation and Rating of Essay Tests.

Educational Testing Service: Princeton, New Jersey.

Diederich, P.B. 1974. Measuring Growth in English. National Council of Teachers of English: Urbana,Illinois.

Diederich, P.B., J.W. French & S.T. Carlton. 1961. Factors in Judgements of Writing Ability. ETS Research Bulletin 61-15. Educational Testing Service: Princeton, N.J.

van Dijk, T.A. 1977. Text and Context: Explorations in the Semantics and Pragmatics of Discourse. Longman: London.

van Dijk, T.A. 1981. 'Discourse studies and education'. Applied Linguistics 2: 1-26.

Dilworth, C.B., R.W. Reising & D.T. Wolfe. 1978. 'Language structure & thought in written composition: certain relationships'. Research in the Teaching of English 12: 97-106.

Dixon, J. & L. Stratta. 1981. Achievements in Writing at 16+. 1: Staging Points Reached in Narratives Based on Personal Experience. Schools Council: London.

Donlan, D. 1974. 'Teaching writing in the content areas: eleven hypotheses from a teacher survey.' Research in the Teaching of English. 8: 250-262.

Dooling, D.J. & R. Lachman. 1971. 'Effects of comprehension on retention of prose'. Journal of Educational Psychology. 88: 221

Douglas, M., ed. Claremont Reading Conference: 42nd Yearbook. Claremont Graduate School, Claremont: California.

Dudley- Evans, A. 1985. 'Aspects of examination questions and answers' in Robinson, ed.

Dudley- Evans, A. 1986. 'Genre-analysis: an investigation of the introduction and discussion sections of MSc dissertations' in Coulthard, ed.

Ebel, R.L. 1951. 'Estimation of the reliability of ratings' Psychometrica 16: 407-424.

Edelsky, C. 1982. 'Writing in a bilingual program: the relation of L1 and L2 texts' TESOL Quarterly 16: 211-228.

Edgeworth, F.Y. 1888. 'The statistics of examinations' Journal of the

Royal Statistical Society 51: 599-635.

Edgeworth, F.Y. 1890. 'The element of chance in competetive examinations' Journal of the Royal Statistical Society 53: 460-475 & 644-663.

Edmiston, R.W. 1939. 'Examine the examination' Journal of Educational Psychology 30: 126-138.

Elbow, P. 1973. Writing Without Teachers. Oxford University Press: Oxford.

Elbow, P. 1986. 'Closing my eyes as I speak: an argument for ignoring audience'. College English 48: 50-69.

Emig, J. 1971. The Composing Processes of Twelfth Graders. NCTE Research Report 13: Urbana, Illinois.

Emig, J. 1978. 'Eye, hand, brain: some "basics" in the writing process' in Cooper & Odell, eds.

Emig, J. 1980. 'The tacit tradition: the inevitability of a multi-disciplinary approach to writing research' in Freedman & Pringle, eds.

Engelskirchen, A. E. Cottrell & J.W. Oller. 1981. 'A study of the reliability and validity of the Ilyin Oral Interview' in Palmer, Groot & Trosper, eds.

Faigley, L. 1986. 'Competing theories of process: a critique and a proposal'. College English 48: 527-542.

Farhady, H. 1979. 'The disjunctive fallacy between discrete-point and integrative tests'. TESOL Quarterly 13: 247-258.

Farhady, H. 1981. Justification, Development and Validation of Functional Language Testing. PhD, University of California at Los Angeles.

Farhady, H. 1982. 'Measures of language proficiency from the learner's perspective'. TESOL Quarterly 16: 46-61.

Fein, D.M. 1980. A Comparison of English and ESL Compositions. MA, University of California at Los Angeles.

Fellows, J.E. 1932. 'The influence of theme-reading and theme-correction

on eliminating technical errors in the written compositions of ninth grade pupils'. Studies in Education 7: 1.

Fillenbaum, S. 1974. 'Pragmatic normalization: further results for some conjunctive and disjunctive sentences'. Journal of Exploratory Psychology 102: 574-508.

Finlayson, D. 1951. 'The reliability of the marking of essays'. British Journal of Educational Psychology 21: 126-134.

Flahive, D. & B.G. Snow. 1980. 'Measures of syntactic complexity in evaluating ESL compositions' in Oller & Perkins, eds.

Flavell, J.H. 1963. The Developmental Psychology of Jean Piaget. Van Nostrand: New York.

Flesch, R. 1974. The Art of Readable Writing. Harper & Row: New York.

Flower, L. 1979. 'Writer-based prose: a cognitive basis for problems in writing'. College English 41: 19-37.

Flower, L. & J. Hayes. 1977. 'Problem-solving strategies and the writing process'. College English 39: 449-461.

Flower, L.S. & J.R. Hayes. 1980. 'The dynamics of composing: making plans and juggling constraints' in Gregg & Steinberg, eds.

Follman, J.C. & J.A. Anderson. 1967. 'An investigation of the reliability of five procedures for grading English themes'. Research in the Teaching of English 1: 190-200.

Foreign Service Institute. no date. F.S.I. Oral Interview Test FSI: Washington, D.C.

Fostvedt, D.R. 1965. 'Criteria for the evaluating of high-school English composition'. Journal of Educational Research 59: 108-112.

Francis, J.C. 1977. Impression & Analytic Marking Methods Mimeo. Associated Examining Board.

Frase, L.T. 1980. 'Writing, text and the reader' in Frederikson et al, eds.

Frederikson, C. & J. Dominic, eds. 1981. Writing: the Nature, Development and Teaching of Written Communications. Vol. 2: Writing: Process, Development & Communication. Lawrence Erlbaum Associates, Hillsdale: N.J.

Freedle, R.O, ed. 1977. Discourse Production and Comprehension Ablex: Norwood, N.J.

Freedle, R.O., ed. New Directions in Discourse Processing. Ablex: Norwood, N.J.

Freedman, A. & I. Pringle, eds. 1980a. Reinventing the Rhetorical Tradition National Council of Teachers of English: Urbana, Illinois.

Freedman, A. & I. Pringle. 1980b. 'Writing in the college years: some indices of growth'. College Composition and Communication 31,3: 311-324.

Freedman, A. & I. Pringle. 1980c. 'Epilogue: reinventing the rhetorical tradition' in Freedman & Pringle, eds.

Freedman, A, I. Pringle & J. Yalden,eds. 1983. Learning to Write: First Language/Second Language. Longman, London.

Freedman, A.W. 1979. 'How characteristics of student essays influence teachers' evaluations' '. Journal of Educational Research 71: 328-338.

Freedman, A.W. 1977. Influences of the Evaluators of Student Writing PhD, Stanford University.

Freedman, S.W. & R.C. Calfee. 1983. 'Holistic assessment of writing: experimental design and cognitive theory' in Mosenthal et al, eds.

Freire, P. 1970. Pedagogy of the Oppressed. Seabury House: New York.

Galbraith, D. 1980. 'The effect of conflicting goals on writing: a case study'. Visible Language 14: 364-375.

Garrett, H.E. 1958. Statistics in Psychology and Education Longman: London.

Gere, A.R. 1981. 'A cultural perspective on talking & writing' in Kroll & Vann, eds.

Gere, A.R. 1980. 'Written composition: toward a theory of evaluation'. College English 42: 44-58.

Giannisi, J. 1976.. 'Dialects & composition' in Tate, ed. 275-304.

Givon, T. 1979. Syntax and Semantics Vol. 12: Discourse and Syntax. Academic Press: New York.

Godman, A. 1980. 'Purposeful testing of language competence'. Paper

presented at SEAMEO RELC Seminar, Singapore.

Godshalk, F.I., F. Swineford & W.E. Coffman. 1966. The Measurement of Writing Ability. ETS Research Monograph 6, Educational Testing Service: Princeton, N.J..

Gonzalez, A. 1979. Ethnocentrism and Teaching Writing to Foreign Students ERIC ED 198 734.

Gosling, G.W. 1966. Marking English Compositions. Australian Council for Educational Research: Victoria, Australia.

Grady, M. 1971. 'A conceptual rhetoric of the composition'. College Composition and Communication 22: 348-354.

Graves, D.H. 1973. Children's Writing: Research Directions and Hypotheses Based Upon an Examination of the Writing Processes of 7 Year Old Children PhD, State University of New York at Buffalo.

Graves, D.H. 1983. 'The growth and development of first grade writers' in Freedman et al, eds.

Green, G.M. & J.L. Morgan. 1981. ' Writing ability as a function of the appreciation of differences between oral and written communication' in Frederikson & Dominic, eds.

Greenall, S. & J. Price, eds. 1980. Study Modes and Academic Development of Overseas Students. ELT Documents 108. British Council: London.

Greenberg, K. 1981. The Effects of Variations in Essay Questions on the Writing of CUNY Freshman CUNY Instructional Resource Center: New York.

Greenberg, K.L. 1985. 'Writing tasks and students' writing performance (1) in Kwalick et al, eds.

Greenberg, K., H. Weiner & R. Donovan. 1986. Writing Assessment: Issues and Strategies. Longman: New York.

Gregg, L.W. & E.R. Steinberg. 1980. Cognitive Processes in Writing: An Interdisciplinary Approach. Lawrence Erlbaum Associates: Hillsdale, N.J.

Grice, H.P. 1975. 'Logic and conversation' in Cole & Morgan, eds.

Grierson, H.J.C. 1944. Rhetoric and English Composition Oliver & Boyd: Edinburgh.

Grobe, C.H. 1981. 'Syntactic maturity, mechanics and vocabulary characteristics as predictors of holistic quality ratings'. Research in the Teaching of English 15: 75-85.

Groot, P.M.J. 1975. 'Validation of language tests' in Palmer & Spolsky, eds.

Gue, L. & E. Holdaway. 1973. 'English proficiency tests as predictors of success in graduate studies in education'. Language Learning 23: 89-103.

Guilford, J.P. 1954. Psychometric Methods. McGraw-Hill: New York.

Guilford, J.P. & B. Fruchter. 1978. Fundamental Statistics in Psychology and Education. McGraw-Hill: New York.

Gullikson, H. 1950. Theory of Mental Tests John Wiley: New York.

Gundlach, R. The Composing Process and the Teaching of Writing: a Study of an Idea and its Uses. PhD, Northwestern University.

Gunnarson, B. 1978. 'A look at the content similarities between intelligence, achievement, personality, and language tests' in Oller & Perkins, eds.

Hagen, L. 1971. 'An analysis of transitional devices in student writing'. Research in the Teaching of English 5: 190-201.

Hairston, M. 1982. 'The winds of change: Thomas Kuhn and the revolution in the teaching of writing'. College Composition and Communication, 33: 76-88.

Hake, R. 1973. Composition Theory in Identifying and Evaluating Essay Writing. PhD, University of Chicago.

Hales, L.W. & E. Tokar. 1975. 'The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question'.Journal of Educational Measurement. 12: 115-117.

Halliday, M.A.K. & R. Hasan. 1976. Cohesion in English. Longman: London.

Halsey, A.H. & M. Trow. 1971. The British Academics. Faber: London.

Hamp-Lyons, L. 1985. 'Three windows on writing'. Paper presented at

SATEFL Conference, Edinburgh.

Hamp-Lyons, L. 1986a. 'Writing in a foreign language and rhetorical transfer: influences on evaluators' ratings' in Meara, ed.

Hamp-Lyons, L. 1986b. 'Proficiency, profiling and M2'. Paper presented at ELTSVAL Conference, British Council, London.

Hamp-Lyons, L. 1986c. 'Testing writing across the curriculum'. Papers in Applied Linguistics, Michigan, 2: 32-48.

Hamp-Lyons, L. 1986d. 'No new lamps for old yet, please'. TESOL Quarterly 20: 790-796.

Hamp-Lyons, L. Forthcoming. 'Postgraduate students and writing: what are the rules of the game?'

Hamp-Lyons, L. & K. Berry Courter. 1984. Research Matters. Newbury House: Rowley, Mass.

Hamp-Lyons, L. & B. Heasley. 1986. Study Writing. Cambridge University Press: London.

Harpin, W. 1976. The Second 'R': Writing Development in the Junior School. George Allen & Unwin: London.

Harris, D. 1969. Testing English as a Second Language. McGraw-Hill: New York.

Harris, R.J. 1962. An Experimental Inquiry into the Functions and Value of Formal Grammar in the Teaching of English, with special reference to the teaching of correct written English to children aged twelve to fourteen. PhD, University of London.

Harris, W.H. 1977. 'Teacher response to student writing: a study of the response patterns of high school English teachers to determine the basis for teacher judgement of student writing'. Research in the Teaching of English, 11: 175-185.

Hartnett, C. 1980. Cohesion as a Teachable Measure of Writing Competence. PhD, University of Pennsylvania.

Hartog, P.J. 1936. 'English Compositions at the School Certificate Examination; and the "Write Anything About Something for Anybody" Theory' in Sadler et al, eds).

Hartog, P.J., P.B. Ballard, P. Gurrey, H.R. Hamley & C. Ebblewhite Smith.

1941. The Marking of English Essays. Macmillan: London.

Hartog, P.J. & E.C. Rhodes. 1935. An Examination of Examinations. Macmillan: London.

Hartwell, P. 1980. 'Dialect interference in writing: a critical view'. Research in the Teaching of English, 14: 101-118.

Haswell, R.H. 1983. 'Minimal marking'. College English, 45: 600-604.

Hatch, E. & H. Farhady. 1982. Research Design and Statistics for Applied Linguists. Newbury House: Rowley, Mass.

Hawkey, R. 1982. An Investigation of Inter-relationships Between Cognitive/Affective and Social Factors and Language Learning. A Longitudinal Study of 27 Overseas Students Using English in Connection with Their Training in the United Kingdom. PhD, University of London.

Hayes, J.R. & L.S. Flower. 1980. 'Writing as problem-solving'. Visible Language. 14: 388-399.

Hayes, J.R. & L.S. Flower. 1983. 'Uncovering cognitive processes in writing: an introduction to protocol analysis' in Mosenthal et al, eds.

Heaton, J.B. 1975. Writing English Language Tests. Longman: London.

Heaton, J.B. ed. 1982. Language Testing. Modern English Publications: Hayes, Middlesex.

Heaton, J.B. & A.K. Pugh. 1974. A Study of the Relationship Between Scores Obtained by Overseas Students on a Test of English Proficiency and Their Examination Results in Their University Courses. School of Education Report, University of Leeds.

Helson, H. 1959. 'Adaptation Level Theory' in Kock, ed.

Hendrickson, J.M. 1985. 'The Treatment of Error in Written Work' in Mackay ed.

Henning, G. 1982. 'A comparison of ratings of EFL composition writing'. Paper presented at the TESOL Convention, Honolulu, Hawaii.

Henning, G. & F. Davidson. 1986. 'Scalar analysis of composition ratings'. Paper presented at the Eighth Annual Language Testing Research Colloquium, Monterrey, California.

Herrington, A. 1985a. 'Classrooms as forums for reasoning and writing'. College Composition and Communication 31: 404-413.

Herrington, A. 1985b. 'Writing in academic settings: a study of the contexts in two college chemical engineeering courses'. Research in the Teaching of English 19: 331-336.

Herrington, A. 1986. 'Studying writing in academic contexts: the view from within our classrooms'. Papers in Applied Linguistics, Michigan 2: 48-64.

Heuring, D.L. 1984. 'Revision strategies of ESL writers: five case studies'. Manuscript of paper presented at TESOL Convention, Houston, Texas.

Hildgard, A. & D. Olson. 1982. 'Examining differences in spoken and written language' in Martlew, ed.

Hilgers, T. 1982. 'Experimental control and the writing stimulus: the problem of unequal familiarity with content'. Research in the Teaching of English 16: 381-390.

Hillyer, J., D. Marcotte & T. Martin. 1969. 'Opinionation, vagueness and specifity distinctions: essay traits measured by computer'. American Educational Research Journal 6: 271-286.

Hillocks, G. 1986. Research on Written Composition: New Directions for Teaching. National Council of Teachers of English: Urbana,Illinois.

Hinds, J. 1983. 'Contrastive rhetoric: Japanese and English'. Text 3: 183-195.

Hinds, J. 1979. 'Organizational patterns in discourse' in Givon, ed.

Hirokawa, K. & J. Swales. 1986. 'The effects of modifying the formality level of ESL composition questions'. TESOL Quarterly, 20: 343-345.

Hirsch, E. 1977., The Philosophy of Composition. University of Chicago Press: Chicago.

Hirsch, E. & D.P. Harrington. 1981. 'Measuring the communicative effectiveness of prose' in Fredericksen & Dominic, eds.

Hoetker, J. 1982. 'Essay examination topics and students' writing'. College Composition and Communication, 33: 377-392.

Hoey, M. 1979. Signalling in Discourse. Discourse Monograph 6. English Language Research, University of Birmingham.

Holland, R.M. 1976. 'Piagetian theory and the design of composing assignments'. Arizona English Bulletin, 19: 17-22.

Hoover, M.R. & R.L. Politzer. 1980. 'Bias in composition tests with suggestions for a culturally appropriate assessment technique' in Whiteman, ed.

Horowitz, D. 1986a. Process not product: less than meets the eye'. TESOL Quarterly, 20: 141-144.

Horowitz, D. 1986b. 'What professors actually require: academic tasks for the ESL classroom'. TESOL Quarterly, 20: 445-462.

Horowitz, D. 1986c. 'The author responds to Hamp-Lyons'. TESOL Quarterly 20: 796-797.

Horowitz, D. Forthcoming.. 'Essay examination prompts and the teaching of academic writing'. Mimeo.

Houghton, D. 1980. 'Contrastive rhetoric'. English Language Research Journal, 1: 79-91.

Houghton ,D. 1984. 'Overseas students writing essays in English: learning the rules of the game' in James, ed.

Houston, J.G. 1983. 'Norm and criterion referencing of performance levels in tests of educational attainment'. Mimeo: Associated Examining Board Advisory Committee (Paper 241).

Howe, P. 1983. Writing Examination Answers. Collins: London.

Howie, D. 1954. 'A comparison of two methods of factorizing test data'. British Journal of Statistical Psychology, 7: 31-36.

Huddleston, E. 1954. 'Measurement of writing ability at the college levels: objective versus subjective testing techniques'. Journal of Experimental Education, 2: 165-213.

Huddleston, E. 1925. 'The effects of objective standards upon composition teachers'. Journal of Educational Research, 12: 329-340.

Hughes, A. 1981a. 'Reaction to the Palmer & Bachman and Vollmer papers' in Alderson & Hughes, eds.

Hughes, A. 1981b. 'Epilogue' in Alderson & Hughes, eds.

Hughes, A. 1981c. 'The structure of language proficiency' in Crystal, ed.

Hughes, A. 1983. 'Prepared comments by Arthur Hughes' in Johnson & Porter, eds.

Hughes, A. & D. Porter. 1983. Current Developments in Language Testing. Academic Press: London.

Hughes, A. & A. Woods. 1981. 'Unitary competence and Cambridge Proficiency'. AILA: Proceedings of 1981 Congress.

Hughes, D.C., B. Keeling & B.F. Tuck. 1983. 'Effects of achievement expectations and handwriting quality on scoring essays'. Journal of Educational Measurement, 20: 65-70.

Hunt, K.W. 1965. 'Grammatical structures written at three grade levels'. Research Report No.3. National Council of Teachers of English: Urbana,Illinois.

Hunt, K.W. 1970. 'Syntactic maturity in school children and adults'. Monographs of the Society for Research in Child Development, 35:134.

Hunt, K.W. 1977. 'Early blooming and late blooming syntactic structures' in Cooper & Odell, eds.

Hymes, D. 1979 'Educational ethnology'. Anthroplogy and Education Quarterly, 11: 31-35.

Ingram, E. 1968. English Language Battery. Mimeo. Dept. Applied Linguistics, University of Edinburgh.

Ingram, E. 1970. 'Report on the marking of English compositions'. Mimeo. Dept. of Linguistics, University of Edinburgh.

Ingram, E. 1973. 'English standards of foreign students'. Edinburgh University Bulletin, May.

Ingram, E. 1977. 'Basic concepts in testing' in Allen & Davies, eds.

Jackson, B. 1965. English Versus Examinations. Chatto & Windus: London.

Jacobs, H.L., S.A. Zinkgraf, D.R. Wormuth, V.F. Hartfiel & J.B. Hughey. 1982. Testing ESL Composition: A Practical Approach. Newbury House: Rowley, Mass.

Jacobs, S. 1982. Composing and Coherence. Center for Applied Linguistics: Washington, D.C.

Jakobovitz, L.A. 1970. Foreign Language Learning: a psycholinguistic analysis f the issues. Newbury House: Rowley, Mass.

James, C. 1980. Contrastive Analysis. Longman: London.

James, G. 1984. The ESP Classroom - Methodology, Materials, Expectations. Vol.7, Exeter Linguistic Studies. University of Exeter, Exeter.

James, K. 'The writing of theses by speakers of English as a foreign language' in Williams et al, eds.

Jeffrey, A.W. ed. 1978. Issues in Educational Measurement. H.M.S.O.: Edinburgh.

Jerabek, R. & P.B. Diederich. 1975. 'Composition evaluation: the state of the art'. College Composition and Communication 26: 183-186.

Johns, A M. 1981. 'Necessary English: an academic survey'. TESOL Quarterly 15: 51-57.

Johns, A. 1985. 'Examining coherence problems in ESL writing'. Paper presented at TESOL Convention, New York.

Johns, T.F. & A. Dudley-Evans. 1980. 'An experiment in team-teaching of overseas postgraduate students of transportation and plant biology'. ELT Documents: British Council.

Johns, C. Lewis. 1981. 'Testing communicative competence'. Midlands Applied Linguistic Studies Journal.

Johnson, D. 1985. 'Error gravity: communicative effect of language errors in academic writing'. Paper presented at SELMOUS Meeting, Reading.

Johnson, K. 1983. 'Communicative writing practice and Aristotelian rhetoric' in Freedman et al, eds.

Johnson, K. & D. Porter eds. 1983. Perspectives in Communicative Language Teaching. Academic Press: London.

Johnstone, A., T.I. Morrison & D.W.A. Sharp. 1971. 'Topic difficulties in Chemistry'. Education in Chemistry 8: 211-218.

Jones, R. 1977. 'Testing: a vital connection' in Phillips, ed.

Jones, R. 1975. 'Achieving objectivity in subjective language tests'. AILA: Proceedings of the 1973 Congress.

Jones, R. & B. Spolsky (eds.). 1975. Testing Language Proficiency. Center for Applied Linguistics: Washington, D.C.

Jones, S. 1986. 'An evaluation scale for integrative writing tests on a performance test of EAP'. Paper presented at Eighth Language Testing Research Colloquium, Monterrey, California.

Jones, T. 1978 'The Foundation course in Laboratory Procedures at King Faisal University CSE Project' ELT Documents: British Council.

Jordan, R.R. 1977. 'Identification of problems and needs: a student profile' in Cowie & Heaton, eds.

Jordan, R.R. & R. Mackay. 1973. 'A survey of the spoken English problems of overseas postgraduates at the universities of Manchester and Newcastle upon Tyne'. Journal (Durham and Newcastle upon Tyne Institutes of Education).

Jordan, R.R. & A. Matthews. 1978. 'English for Specific Purposes: practice material for the listening comprehension and writing needs of overseas students'. ELT Documents: The British Council.

Jurgens, J.M. & W.J. Griffin. 1970. 'Relationships between overall quality and seven language features in grades seven, nine and eleven'. ERIC ED 046 932.

Kameen, P. 1983. 'Syntactic skill and ESL writing quality' in Freedman et al, eds.

Kaplan, R.B. 1966. 'Cultural thought patterns in intercultural education'. Language Learning 16: 1-20.

Kaplan, R.B. 1967. 'Contrastive rhetoric and the teaching of composition'. TESOL Quarterly 1: 10-16.

Kaplan, R.B. 1972. The Anatomy of Rhetoric: Prolegomena to a Functional Theory of Rhetoric. Center for Curriculum Development, Inc.: Philadelphia, PA.

Kaplan R.B., ed. 1981. Annual Review of Applied Linguistics. Vol.1. Newbury House: Rowley, Mass.

Kaplan R.B. 1983. 'Contrastive rhetorics: some implications for the writing process' in Freedman et al, eds.

Kaplan, R.B., ed. 1983. Annual Review of Applied Linguistics, Vol.3. Newbury House: Rowley, Mass.

Kelly, L.G. 1969. 25 Centuries of Language Teaching. Newbury House: Rowley, Mass.

Kempson, R. 1975. Presupposition and the Delimitation of Semantics. Cambridge University Press: London.

Kiefer, K. & C. Smith. 1984. 'Improving students' revising and editing: the Writer's Workbench system' in Wresch, ed.

Kincaid, G.L. 1953. Some Factors Affecting Variations in the Quality of Students' Writing. PhD, Michigan State University.

Kinneavy, J.L. 1980. 'A pluralist synthesis of four contemporary models for teaching composition' in Freedman & Pringle, eds.

Klare, D., ed. 1976. Cognition and Instruction. Lawrence Erlbaum Associates: New Jersey.

Klein-Braley, C. 1985. 'A cloze-up on the C-test: a study in the construct validation of authentic tests'. Language Testing 2: 76-104.

Klein Braley, C. & D. Stevenson. 1981. Orbis Linguisticus Vol.1: Practice and Problems in Language Testing 1. Verlag Peter D Lang: Frankfurt.

Knorr-Cetina, K. 1981. The Manufacture of Knowledge. Pergamon Press: Oxford.

Kock, S (ed.) 1959. Psychology: A Study of a Science. McGraw-Hill: New York.

Kojima, S. & K. Kojima. 1978. 'S (Inanimate Subject) + V + O: A syntactical problem in EST writing for Japanese' in Todd Trimble, Trimble & Drobnic, eds.

Kolb, D. 1981. 'Learning styles and disciplinary differences' in Chickering, ed.

Kozminsky, E. 1977. 'Altering comprehension: the effect of biasing titles on text comprehension'. Memory & Cognition 5: 482-490.

Krashen, S. 1978. 'On the acquisition of unplanned discourse: written English as a second dialect' in Douglas, ed.

Kroll, B. 1979. 'A survey of the writing needs of foreign and American college freshmen'. English Language Teaching Journal 33: 219-226.

Kroll, B. 1982. Levels of Error in ESL Composition. PhD, University of Southern California.

Kroll, B.M. & J.C. Schafer. 1985. 'Error analysis and the teaching of composition' in Mackay, ed.

Kroll, B.M. & R. Vann. 1981. Exploring Speaking-Writing Relationships: Connections and Contrasts. National Council of Teachers of English: Urbana,Illinois.

Kuder, G.F. & M.W. Richardson. 1937. 'The theory of the estimation of test reliability'. Psychometrika 2: 151-160.

Kuhn, T. 1970. The Structure of Scientific Revolutions. University of Chicago Press: Chicago.

Kwalick, B., M. Silver & V.B. Slaughter. 1983. Selected Papers from the 1982 Conference: New York Writes. CUNY Instructional Resource Center: New York.

Labov, W 1969. 'The logic of non-standard English' in Alatis, ed.

Labov, W. & D. Fanshel. 1977. Therapeutic Discourse. Academic Press: New York.

Lacey, C. & D. Lawton, eds. 1981. Issues in Evaluation and Accountability. Methuen: London.

Lado, R. 1961. Language Testing: the construction and use of foreign language tests. Longman: London.

Land, G. 1983. 'A made-to-measure ESP course for banking staff'. ESP Journal 2: 161-171.

Larsen-Freeman, D. 1978. 'An ESL index of development'. TESOL Quarterly 12: 439-448.

Larson, R. 1968. 'Discovery through questioning: a plan for teaching rhetorical invention'. College English 30: 126-134.

Larson, R. 1971. 'Toward a linear rhetoric of the essay'. College Composition and Communication 22: 14-145.

Lay, N. 1982. 'Composing processes of adult ESL learners: a case study'. TESOL Quarterly 16: 406.

Lay, N. 1983. 'Native language and the composing process' in Kwalick et al, eds.

Lees, E.O. 1979. 'Evaluating student writing'. College Composition and Communication 30: 370-374.

Li, C. & S. Thompson. 1976. Subject and Topic. Academic Press, New York.

Light, D. Jr. 'The structure of the academic professions'. Sociology of Education 47: 2-28.

Lim, Ho-Peng. 1984. 'Measuring writing proficiency of college ESL students'. JALT Journal. 6: 12-14.

Lindstrom, M.W. 1981. Native Speaker Reactions to Stylistic Errors in Writing: an Error Evaluation. MA, Colorado State University.

Lloyd-Jones, R. 1978. 'Primary trait scoring' in Cooper & Odell, eds.

Low, G. 1982. 'Direct writing tests'. SYSTEM 10: 132-137.

Lubin, G., J. Margary & M. Poulsen. eds. 1975. Piagetian Theory and Its Implications for the Helping Professions. University of Southern California, Los Angeles.

Lucas, A.M. 1971. 'Multiple marking of a matriculation Biology essay question'. British Journal of Educational Psychology 41,1.

Lunsford, A. 1986. 'The past - and future - of writing assessment' in Greenberg et al, eds.

Luria, A.R. 1976. Cognitive Development. Its Cultural and Social Foundation. Harvard University Press, Harvard.

Lynch, A. 1983. Study Listening. Cambridge University Press: London.

Mackay, R. & A. Mountford, eds. 1978a. English for Specific Purposes. Longman: London.

Mackay, R & A. Mountford. 1978b. 'The teaching of English for special purposes: theory and practice' in Mackay & Mountford, eds.

Mackay, S. 1985. Composing in a Second Language. Newbury House: Rowley, Mass.

Madson, H.S. & R. Jones. 1981. 'Classification of oral proficiency tests' in Groot, Palmer & Trosper, eds.

Maher, J. 1985. The Role of English in Medicine and Medical Education in Japan. PhD, University of Edinburgh.

Markham, L. 1976. 'Influences of handwriting quality on teacher evaluation of written work'. American Educational Research Journal 13, 277-283.

Marsden, R. & N. Underhill. 1981. 'A comparison of some common types of language tests'. Paper presented at AILA Congress, Brussels.

Martlew, M. 1983. The Psychology of Writing: development and educational perspectives. John Wiley: London.

Marton, F., D.J. Hounsell & N.J. Entwhistle . 1984. The Experience of Learning. Scottish Academic Press: Edinburgh.

Marton, F. & R. Saljo. 1976. 'On qualitative differences in learning: 1 - Outcome and Process'. British Journal of Educational Psychology 46: 4-11.

Maxwell, A.E. 1977. Multivariate Analysis in Behavioural Research. Chapman & Hall: London.

Maxwell, J.C. 1973. 'National assessment of writing: useless and uninteresting? English Journal 62: 1254-1257.

McColly, W. 1970. 'What does educational research say about the judging of writing ability? Journal of Educational Research 64: 148-156.

McColly, W. & R. Romstad. 1965. 'Composition rating scales for general merit: an experimental evaluation'. Journal of Educational Research 59: 55-56.

McCully, B. 1965. English Education and the Origins of Indian Nationalism. Columbia University Press: New York.

McEldowney, P. 1976. 'Test in English (Overseas). The position after ten years.' JMB Occasional Papers 36: University of Manchester.

McGinley, K. 1983. 'Some notes on evaluation in ESP>. Paper presented at the SELMOUS Meeting, Exeter.

McKenna, E., F. Clark & F. Zorn. Forthcoming. 'Studying the longitudinal study'.

McQuade, D, ed. 1979. Linguistics, Stylistics and the Teaching of Composition. University of Akron Press, Akron, OH.

Mead, R. 1978 'Student needs and the authenticity of ESP materials', MALS Journal.

Meara, P. ed. 1986. Spoken Language. BAAL/CILT: London.

Mendelsohn, D.J. & M. Tyacke. 1981. 'Rapid assessment of writing for placement: a comparison of six methods'. Mimeo.

Meredith, V.H. & P.L. Williams. 1984. 'Issues in direct writing assessment: problem identification and control'. Educational Measurement: Issues and Practices 3: 12-15.

Messick, S.A. 1975. 'The standard problem: meaning and values in measurement and evaluation'. American Psychologist 30: 955-966

Messick, S.A. 1980. 'Test validity and the effects of assessment'. American Psychologist 35: 1012-1027.

Meyer, B.J.F. 1975. The Organization of Prose and Its Effect on Recall. North Holland: New York.

Meyer, B.J.F. 1977. 'What is remembered from prose: a function of passage structure' in Freedle, ed.

Meyer, B.J.F. 1985. 'Reading research and the composition teacher: the importance of plans' in Mackay, ed.

Meyer, B.J.F. & G.E. Rice. 1982. 'The interaction of reader strategies and the organization of text'. Text 2: 155-192.

Meyer, G. 1935a. 'An experimental study of the old and new types of examination. I: The effect of the examination set on memory'. Journal of Educational Psychology 25: 641-661.

Meyer, G. 1935b. II: Methods of study'. Journal of Educational Psychology 26: 30-40.

Meyer, G. 1939. 'The choice of questions on essay examinations'. Journal of Educational Psychology 30: 161-171.

Meyers, A., C. McConville & W. Coffman. 1966. 'Simplex structures in the grading of essay tests'. Educational and Psychological Measurement 2: 41-45.

Miller, J. & W. Kintsch. 1980. 'Readability and recall of short, prose passages: A theoretical analysis' Journal of Educational Psychology: Human Learning and Memory 6: 335-354.

Miller, S. Passler. 1976. Writing: Process and Product. Winthrop: Cambridge.

Miller, S. Passler. 1982. 'How writers evaluate their own writing'. College Composition and Communication33: 176-183.

Mitchell, R., B. Parkinson & R. Johnstone. 1982. The Foreign Language Classroom: An Observational Study. Stirling Education Monographs, No.9. University of Stirling.

Mitler, W, ed. 1979. The Use of Tests and Interviews for Admission to Higher Education. Council of Europe/NFER: Windsor, Berks.

Mohan, B.A. & W. Au-Yeung Lo. 1986. 'Academic writing and Chinese students: transfer and developmental factors'. TESOL Quarterly 19: 515-534.

Moller, A. 1980. 'Assessing Proficiency in English for use in further study'. Paper presented at RELC Seminar, Singapore.

Moller, A. 1981. 'Reaction to the Morrow paper' in Alderson & Hughes, eds.

Moller, A. 1982. A Study in the Validation of Proficiency Tests of English as a Foreign Language. PhD, University of Edinburgh.

Morris, B.S. 19676. International Community? National Union of Students of England, Wales, N. Ireland, and Scottish Union of Students: London.

Morris, L. 1985. 'Classroom research in writing assessment'. NTNW Notes. Nov. 1985.

Morrison, J.W. 1974. An Investigation of Problems in Listening Comprehension Encountered by Overseas Students in the First Year of Postgraduate Studies in Sciences in the University of Newcastle upon Tyne and the Implications for Teaching. MEd, University of Newcastle upon Tyne.

Morrison, R.L. & P.E. Vernon. 1941. 'A new method of marking English

compositions'. British Journal of Educational Psychology 11: 109-119.

Morrow, K. 1977. Techniques of evaluation for a notional syllabus. Royal Society of Arts: London.

Morrow, K. 1981. 'Communicative language testing: revolution or evolution?' in Alderson & Hughes, eds.

Mosenthal, P., L. Tamor & S. Walmesley, eds. 1983. Research in Writing: Principles and Methods. Longman: New York.

Mullen K. 1980. 'Evaluating writing proficiency in ESL' in Oller & Perkins, eds.

Mullis, I. 1980. Using the Primary Trait System for Evaluating Writing. NAEP: Education Commission of the States, Denver, CO.

Mullis, I. 1984. 'Scoring direct writing assessments: what are the alternatives?' Educational Measurement: Issues and Practice 3: 16-18.

Munby, J. 1977. Designing a Processing Model for Specifying Communicative Competence in a Foreign Language: a study of the relationship between communication needs and the English required for specific purposes. PhD, University of Essex.

Munby, J. 1978. Communicative Syllabus Design. Cambridge University Press: London.

Murakami, M. 1980. 'Behavioural and attitudinal correlates of progress in ESL by native speakers of Japanese' in Oller & Perkins, eds.

Murphy, S. & J. Keroes. 1985. 'Essay test topic development'. NTNW Notes, Nov.1985.

Murray, D.H. 1968. A Teacher Teaches Writing. Houghton Mifflin: Boston.

Murray, D.H. 1978. 'Internal revision: a process of discovery' in Cooper & Odell, eds.

Myers, G. 1985. 'The social construction of two biologists' proposals'. Written Communication 2: 219-245.

National Assessment of Educational Progress. 1980. Writing Achievement 1969-79: Results from the Third National Writing Assessment. NAEP: Education Commission of the States, Denver, CO.

Nelson, L. & G. Piche. 1981. 'The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing'. Research in the Teaching of English 15: 65-73

Newcomb, J.S. 1977. The Influence of Readers on the Holistic Grading of Essays. PhD, University of Michigan.

Newell, A. & H. Swain. 1972. Human Problem-Solving. Prentice-Hall: Englewood Cliffs, N.J.

New York State Education Dept. 1972. 'Sample examination questions. Grade 9 Social Studies'. New York State Education Dept, Albany: Bureau of Social Studies Education.

Nold, E. & B. Davis. 1980. 'The discourse matrix'. College Composition and Communication 31: 141-152.

Nunnally, J.C. 1964. Educational Measurement and Evaluation McGraw-Hill: New York.

Nunnally, J.C 1978. Psychometric Theory. McGraw-Hill: New York.

Nuttall, D. & L. Skurnik. 1969. Examination and Item Analysis Manual NFER· London.

Nystrand, M, ed. 1983a. What Writers Know: The Language, Process and Structure of Written Discourse. Academic Press: London.

Nystrand, M. 1983b. 'An analysis of errors in written communication' in Nystrand, ed.

Nystrand, M. 1983c. 'Rhetoric's "Audience" and linguistics' "Speech Community": implications for understanding reading, writing and text' in Nystrand, ed.

Nystrand, M. 1983d. 'The structure of textual space' in Nystrand, ed.

O'Brien, T. 1985. 'Writing for continuous assessment or examinations - a comparison of style'. Paper presented at SELMOUS Meeting, Reading.

O'Donnell, W. 1968. An Investigation into the Role of Language in a Physics Examination. Moray House Monograph No. 7. Oliver & Boyd: Edinburgh.

Odell, L. 1973. 'Piaget, problem-solving and freshman composition'. College Composition and Communication 24: 36-42.

Odell, L. 1976. 'The classroom teacher as researcher'. English Journal 65: 106-111.

Odell, L. 1981. 'Defining and assessing competence in writing' in Cooper, ed.

Odell, L., C.R. Cooper & C. Courts. 1978. 'Discourse theory: implications for research in composing' in Cooper & Odell, eds.

Oller, J.W. 1971. 'Dictation as a device for testing foreign language proficiency'. English Language Teaching Journal 25: 254-259.

Oller, J.W. 1972. 'Scoring methods and difficulty levels for cloze tests of proficiency in ESL'. Modern Language Journal 56: 151-158.

Oller, J.W 1974. 'Expectancy for successive elements: key ingredient to language use'. Foreign Language Annals 7: 443-452.

Oller, J.W. 1978a. 'Pragmatics and language testing' in Spolsky, ed.

Oller, J W. 1978b. 'How important is language proficiency to IQ and other educational tests?' in Oller & Perkins, eds.

Oller, J.W. 1979a. Language Tests at School. Longman: London.

Oller, J.W. 1979b. 'Explaining the reliable variance in tests: the validation problem' in Briere & Hinofotis, eds.

Oller, J.W. 1980. 'Communicative competence: can it be tested?' in Scarcella & Krashen, eds.

Oller, J.W. 1983. Issues in Language Testing Research. Newbury House: Rowley, Mass.

Oller, J.W. & C. Conrad. 1971. 'The cloze procedure and ESL proficiency'. Language Learning 21: 183-196.

Oller, J.W. & F. Hinofotis. 1980. 'Two mutually exclusive hypotheses about

second language proficiency: indivisible or partially divisible competence' in Oller & Perkins, eds.

Oller, J.W. & F. Khan. 1980. 'Is there a global factor of language proficiency?' Paper presented at RELC Seminar, Singapore.

Oller, J.W. & K. Perkins, eds. 1978. Language in Education: Testing the

Tests. Newbury House: Rowley, Mass.

Oller,J.W. & K. Perkins, eds. 1980. Research in Language Testing. Newbury House: Rowley, Mass.

Oller, J.W. & J. Richards. 1973. Focus on the Learner: pragmatic perspectives for the language teacher. Newbury House: Rowley, Mass.

Ostler, S. 1980. 'A survey of academic needs for advanced ESL'. TESOL Quarterly 14: 489-502.

Ong, W. 1975. 'The writer's audience is always a fiction'. PMLA 90: 9-21.

Page, E. 1967. 'Grading essays by computer' in Proceedings of the 1966 Invitational Conference on Testing Procedures, Educational Testing Service: Princeton, N.J.

Palmer, A. & L. Bachman. 1981. 'Basic concerns in test validation' in Alderson & Hughes, eds.

Palmer, A., P. Groot & G. Trosper. 1981. The Construct Validation of Tests of Communicative Competence. TESOL: Washington, D.C.

Palmer, A. & M. Kimball. n.d. A criterion-based composition grading system. Mimeo.

Palmer, A. & B. Spolsky, eds. Papers on Language Testing. TESOL: Washington, D.C.

Parkinson, B., R. Mitchell & R. Johnstone. 1981. Mastery Learning in Foreign Language Teaching: a case study. Stirling Education Monographs No. 8: University of Stirling.

Park, D. 1986. 'Analyzing audiences'. College Composition and Communication 37: 478-488.

Peel, E.A. 1971. The Nature of Adolescent Judgement. Staples Press: London.

Pellegrini, A. & T. Yawkey, eds. 1984. Development of Oral and Written Language. Ablex: Norwood, N.J.

Penfold, E.D.M. 'Essay marking experiments: shorter and longer essays'. British Journal of Educational Psychology 26: 129-136.

Perelman, C. & L. Ulbrechts-Tyteca. 1969. The New Rhetoric: A Treatise on Argumentation. University of Notre Dame Press, South Bend, IN.

Perkins, K. 1980. 'Using objective methods of attained writing proficiency to discriminate among holistic evaluations'. TESOL Quarterly 14: 61-70.

Perkins, K. & R. Leahy. 1979. 'Using objective measures of composition to compare native and non-native compositions'. Paper presented at Third International Symposium on Language Proficiency and Dominance Testing, Carbondale, IL.

Perl, S. 1979. 'The composing processes of unskilled college writers'. Research in the Teaching of English 13: 317-336.

Perl, S. 1981. Coding the composing process: a guide for teachers and researchers. Mimeo. National Institute Education: Washington, D.C.

Perren, G.E. & J. Trim, eds. 1971. Applications of Linguistics. AILA: Proceedings of the 1969 Congress.

Perron, J.D. 1977. 'The impact of mode on written syntactic complexity'. Parts I - III. Studies in Language Education: Reports 24, 25, 27. University of Georgia: Athens, GA.

Phillips, D., E. Burke, A. Campbell & D. Ingram. 1985. The Formative Evaluation of Preparatory English Language Training of Sponsored Indonesian Students. Australian Development Assistance Bureau: Canberra.

Phillips, J.K, ed 1977. The Language Connection. ACTFL/National Textbook Company: Evanston, IL.

Pianko, S. 1978. 'A description of the composing processes of college freshman writers'. Research in the Teaching of English 13: 5-22.

Picus, M. 1984. 'When Asians write: what to expect in rhetoric'. TECFORS 7,4: 11-15.

Pike, L.W. 1973. An evaluation of alternative item formats for testing English as a Foreign Language. ETS Research Report 79-6: Educational Testing Service: Princeton, N.J.

Pilliner, A.E.G. 1968. 'Subjective and objective testing' in Davies, ed.

Pilliner, A.E.G. 1969. 'Multiple marking: Wiseman or Cox?' British

Journal of Educational Psychology 39: 313-315.

Pilliner, A E.G. 1978 'Norm-referenced and criterion-referenced tests - an evaluation' in Jeffrey, ed.

Pilliner, A.E.G. 1982. 'Evaluation' in Heaton, ed.

Pimsleur, P. 1968. 'Language aptitude testing' in Davies, ed.

Pimsleur, P. & T. Quinn. 1971. The Psychology of Second Language Learning. AILA: Proceedings of the 1969 AILA Congress.

Pitcher, B. & J.B. Ra. 1967. The Relation Between Scores on the TOEFL and Ratings of Actual Theme Writing. ETS Statistical Report 67-9. Educational Testing Service: Princeton, N.J.

Poetker, J.S. 1976. 'Constructing better essay questions'. Social Studies Journal 5: 71-74.

Poetker, J.S. 1977. 'Practical suggestions for improving and using essay questions'. High School Journal 61: 7-15.

Politzer, R.L. & M.R. Hoover. 1977. Attitudes Toward Black English Speech Varieties and Black Pupils' Achievement. Stanford Center for Research and Development in Teaching: Stanford, CA.

Pollitt, A. 1978. 'Item-banking' in Jeffrey, ed.

Pollitt, A. & C. Hutchinson. no date. The Dunning Credit Level Project Subject Report: English. Godfrey Thompson Research Unit: Edinburgh.

Pollitt, A. & C. Hutchinson. no date. TELS Mimeo. Godfrey Thompson Research Unit: Edinburgh.

Pollitt, A., C. Hutchinson, N, Entwhistle & C. Deluca. 1985. What Makes Exam Questions Difficult? An Analysis of 'O' Grade Questions and Answers. Research Reports for Teachers No. 2. Scottish Academic Press: Edinburgh.

Popham, W.J. 1975. Educational Evaluation. Prentice-Hall: Englewood Cliffs, N.J.

Popham, W.J. 1978. Criterion-Referenced Measurement. Prentice-Hall: Englewood Cliffs, N.J.

Polanyi, M. 1958. Personal Knowledge: Towards a Post-Critical Philosophy. University of Chicago Press: Chicago, IL.

Assignments on the Performance in English Composition of a Selected Group of 15/16 Year Old Pupils. PhD, University of London.

Rounds, P. 1984. A discourse analysis approach to summary writing. Mimeo.

Rubin, A. 1980. 'A theoretical taxonomy of the differences between oral and written language' in Spiro, Brice & Brewer, eds.

Rubin, D.L. 1982. 'Adapting syntax in writing to varying audiences as a function of age and social cognitive ability'. Journal of Child Language 9: 497-510.

Rubin, D.L. & G.L. Pichè. 1979. 'Development in syntactic and strategic aspects of audience adaptation skills in written persuasive communication'. Research in the Teaching of English 18: 293-316.

Rumelhart, D.E. 1975. 'Notes for a schema for stories' in Bobrow & Collins, eds.

Ruth, L. 1982. Properties of Writing Tasks: A study of alternative procedures for holistic writing assessment. Final Report NIE G-80-0034. Bay Area Writing Project, University of California, Berkeley

Ruth, L. & S. Murphy. 1984. 'Designing topics for writing assessment'. College Composition and Communication 35: 410-422.

Ruth, L. & S. Murphy. Forthcoming. Designing Writing Tasks for the Assessment of Writing. Ablex: Norwood, N.J.

Sachse, P.P. 'Writing assessment in Texas: practices and problems'. Educational Measurement: Issues and Practice 3: 21-23.

Sadler, M., A. Abbott, P.B. Ballard, C.L. Burt, C.D. Burns, P. Hartog, C. Spearman & J. D. Stirk. 1936. Essays on Examinations. Macmillan: London.

Salimbene, S. 1985. Strengthen Your Study Skills! Newbury House: Rowley, Mass.

Sanders, S. & J. Littlefield. 1975. 'Perhaps test essays can reflect siginificant improvement in freshman composition'. Research in the Teaching of English 9: 145-153.

Sang, F., B. Schmidt, H.J. Vollmer, J. Baumert & P.M. Roeder. 1986. 'Models of second language competence: a structural equation approach'. Language Testing 3: 54-79.

Scaglione, A. 1972. The Classical Theory of Composition. University of North Carolina Press: Chapel Hill, NC.

Scarcella, R. 1984. Cohesion in the Writing Development of Native and Non-Native English Speakers. PhD, University of Southern California.

Scardamalia, M. 1975. 'Two formal operational tasks: a quantitative neo-Piagetian and task analysis model for investigating sources of task difficulty' in Lubin, Magary & Poulsen, eds.

Scardamalia, M., C. Bereiter & H. Goelman. 1983. 'The role of production factors in writing ability' in Nystrand, ed.

Schank, R. & R. Abelson. 1977. Scripts, Plans, Goals and Understanding: An inquiry into human knowledge structures. Lawrence Erlbaum Associates: Hillsdale, N.J.

Schneider, M. 1984. 'Judging writing quality: does cohesion make a difference?' Paper presented at TESOL Convention, Houston, TX.

Schroder, H.M., H.J. Driver & S. Streufert. 1967. Human Information Processing. Holt, Rinehart & Winston: New York.

Schwartz, M. 1984. 'Response to writing: a college-wide perspective'. College English 46: 55-62.

Searle, J.R. 1975. 'Indirect speech acts' in Cole & Morgan, eds.

Seaton, I. 1980. 'An English Language Testing Service: subject/language colloboration in ESP test design'. ELT Documents: British Council.

Seaton, I. 1981. 'A review of issues raised in the production of the English Language Testing Service' in Klein-Braley & Stevenson, eds.

Seaton, I. 1983. Issues in the development and operation of the English language Testing Service (ELTS) 1976-1983. Mimeo.

Selected Papers from the 1981 Texas Writing Research Conference. University of Texas Dept of English: Austin, TX.

Sen, A. 1970 Problems of Overseas Students and Nurses. NFER: Windsor, Berks.

Selinker, L. & B. Kamaradevelu 1986. 'An interlanguage/safe-rules approach to composing in L2'. Papers in Applied Linguistics, Michigan. 1.2: 1-30.

Shaughnessy, M. 1977. Errors and Expectations, a guide for the teacher of basic writing. Oxford University Press: New York.

Shohamy, E. & T. Reves. 1985. 'Authentic language tests: where from and where to?' Language Testing 2: 48-59.

Shuy, R. 1981. 'Toward a developmental theory of writing' in Frederikson & Dominic, eds.

Shuy, R. & R. Fasold. 1973. Language Attitudes: current trends and prospects. Georgetown University Press: Washington, D.C.

Silverstein, R, ed. 1979. Proceedings of the Third International Conference on Frontiers in Language Proficiency and Dominance Testing. SIU Occasional Papers on Linguistics No. 6: Carbondale, IL.

Sim, D. & B. Laufer-Dvorkin. 1982. Reading Comprehension Course. Collins: London.

Simon, H.A. & J.R. Hayes. 1976. 'Understanding complex task instructions' in Klare, ed.

Sims, V.M. 1933. 'Reducing the variability of essay examination marks through eliminating variations in standards of grading'. Journal of Educational Research 26: 637-647.

Sinclair, J. McH. 1978. 'Issues in current ESP project design and management'. Midlands Applied Linguistics Journal.

Skehan, P. 1984. 'Issues in the testing of English for specific purposes'. Language Testing 1: 202-220.

Slakter, M.J. 1966. Statistical Inference for Educational Research. Addison-Wesley: Reading, Mass.

Slotnick, H.B. 1972. 'Towards a theory of computer essay grading'. Journal of Educational Measurement 9: 253-263.

Slotnick, H.B. 1973. 'On the teaching of writing: some implications from National Assessment'. English Journal 62: 1248-1253.

Smith, F. 1982. Writing and the Writer. Heinemann: London.

Smith, R. 1975. Grading the Advanced Placement English Examination. CEEB: Princeton. N.J.

Smith, V.H. 1969. 'Measuring teacher judgement in the evaluation of written composition'. Research in the Teaching of English 3: 181-195.

Smith, W.L. & M.B. Swan. 1978. 'Adjusting syntactic structures to varied levels of audience'. Journal of Experimental Education 46: 429-434.

Soloff, S. 1973. 'Effect of noncontent factors on the grading of essays'. Graduate Research in Eduvcation and Related Disciplines 6: 44-54.

Sommers, N. 1980. 'Revision strategies of student writers and experienced adult writers'. College Composition and Communication 36: 82-93.

Sommers, N. 1982. 'Responding to student writing'. College Composition and Communication 33: 148-156.

Spack, R. 1984. 'Invention strategies and the ESL college composition student'. TESOL Quarterly 18: 649-670.

Spearman, C. 1936. 'Note on the reliability and validity of measurements' in Sadler et al.

Spearman, C.E. 1904. '"General intelligence" objectively determined and measured'. American Journal of Psychology 15: 201-293.

Spencer, R.E. & P.D. Holtzman. 1965. 'It's composition - but is it reliable?' College Composition and Communication 21: 117-121.

Spiro, R., B. Bruce & W. Brewer, eds. 1980. Theoretical Issues in Reading Comprehension. Lawrence Erlbaum Associates: Hillsdale, N.J.

Spolsky, B. 1973. 'What does it mean to know a language? or how do you get someone to perform his competence? in Oller & Richards, eds.

Spolsky, B. 'Language testing - the problem of validation' in Palmer & Spolsky, eds.

Spolsky, B, ed. 1978a. Approaches to Language Testing. Vol.2 in Advance in Language Testing Series. Center for Applied Linguistics: Washinton, D.C.

Spolsky, B. 1978b. 'Language testing: art or science?' in Spolsky, ed.

Spolsky, B. 1981. 'Some ethical questions about language testing' in Klein-Braley & Stevenson, eds.

Spolsky, B. 1985. 'The limits of authenticity in language testing'. Language Testing 2: 31-40.

Stallard, C.K. 1974. 'An analysis of the writing behaviour of good student writers'. Research in the Teaching of English 8: 206-218.

Stalnaker, J.M. 1934. 'The construction and results of a twelve-hour test in English composition'. School and Society 39: 218-224.

Stalnaker, J.M. 1937. 'Essay examinations reliably read'. School and Society 56: 671-672.

Stalnaker, J.M. & R.C. Stalnaker. 1934. 'Reliable reading of essay tests'. School Review 52: 599-605.

Stansfield, C, ed. 1986. Technology and Language Testing. TESOL: Washington, D.C.

Stansfield, C. & R. Webster. 1986. The New TOEFL Writing Test. ETS Publicity Mimeo: Educational Testing Service: Princeton, N.J.

Starch, D. & E C. Elliott. 1912. 'Reliability of grading high school work in English'. School Review 20: 442-457.

Stenner, A.J., M. Smith & D.S. Burdick. 1983. 'Toward a theory of construct definition'. Journal of Educational Measurement 20: 305 316.

Sternberg, R. 1979. 'Stalking the IQ quark'. Psychology Today 13: 42-54.

Sternglass, M. 1977. 'Applications of the Wilkinson model of writing maturity to college writing'. College Composition and Communication 33: 167-175.

Stevenson, D.K. 1975. A Preliminary Investigation of Construct Validity and the Test of English as a Foreign Language. PhD, University of New Mexico.

Stevenson, D.K. 1981a. 'Language testing and academic accountability: on redefining the role of language testing in language teaching'. International Review of Applied Linguistics 19: 15-30.

Stevenson, D.K. 1981b. 'Beyond faith and face validity: the multitrait - multimethod matrix and the convergent and discriminant validity of oral proficiency tests' in Palmer, Groot & Trosper, eds.

Stevenson, D.K. 1985. 'Authenticity, validity and a tea party'. Language Testing 2: 41-47.

Stevenson, D.K. & U. Riewe. 1982. 'Teachers' attitudes towards language tests and testing' in Klein-Braley & Stevenson, eds.

Stewart, M.A. 1978. 'Syntactic maturity from high school to college: a first look'. Research in the Teaching of English 12: 37-46.

Stewart, M.F. 1983. 'Teachers' writing assessments across the high school curriculum'. Research in the Teaching of English 17: 113-125.

Stiggins, R. 1984. 'Scoring procedures: holistic, analytic and primary trait'. NTNW Notes, May 1984.

Strevens, P. 1977a. New Orientations in the Teaching of English. Oxford University Press: Oxford.

Strevens, P. 1977b. 'Special-purpose language teaching: a perspective'. Language Teaching and Linguistics Abstracts 10.

Stump, T.A. 1978. 'Cloze and dictation tasks as predictors of intelligence and achievement' in Oller & Perkins, eds.

Sutherland, P.A.A. 1982. 'An expansion of Peel's describer-explainer stage theory'. Education Review 34: 69-76.

Swales, J. 1981. Aspects of Article Introductions. Mimeo, University of Aston, Birmingham.

Swales, J. 1982. 'Examining examination papers'. English Language Research Journal 3: 9-25.

Swales, J. 1984a. 'Research into the structure of introductions to journal articles and its application to the teaching of academic writing' in Williams et al, eds.

Swales, J. 1984b. 'Thoughts on, in and outside the classroom' in James, ed. Swales, J. 1986a. Episodes in ESP. Pergamon Press: Oxford.

Swales, J. 1986b. Utilizing the literatures in teaching the research paper. Mimeo.

Swales, J. n.d. Non-native speaker graduate engineering students and their introductions: global coherence and local management. Mimeo.

Swartz, H., L.S. Flower & J.R. Hayes. 1980. How headings in documents can mislead readers. Technical Report No.9., Document Design Project. Carnegie-Mellon University: Washington, D.C.

Swartz, H., L.S. Flower & J.R. Hayes. 1984. 'Designing protocol studies of the writing process: an introduction' in Bridwell & Beach, eds.

Sweedler-Brown, C.O. 1985. 'The influence of training and experience on holistic essay evaluations'. English Journal 74: 49-55.

Sweet, H. 1971. The Principles of Spelling Reform. (First published 1877; reprinted in The Indispensable Foundation, ed. by E.J.A. Henderson.)

Swineford, F. 1956. College Entrance Examination Board General Composition Test, Form L. Statistical Report 56-45. ETS: Princeton, N.J.

Tamor, L. & J.T. Bond. 1983. 'Text analysis: inferring process from product' in Mosenthal et al, eds.

Tannen, D. 1979. 'What's in a frame? Surface evidence for underlying expectations.' in Freedle, ed.

Tannen, D. 1980. 'A comparative analysis of oral narrative strategies: Athenian Greek and American English' in Chafe, ed.

Tarone, E., U. Frauenfelder & L. Selinker. 1976. 'Systematicity/variability and stability/instability in interlanguage systems' in Brown, ed.

Tate, G, ed. 1976. Teaching Composition: Ten Bibliographical Essays. Texas Christian University Press: Fort Worth, TX.

Taunton, H.L. (Chairman) 1868. Report of the Schools Inquiry Commission.

Eyre & Spottiswoode: London.

Taylor, B. 1981. 'Content and written form: a two-way street'. TESOL Quarterly 15: 5-13.

Thomas, D. & Donlan, D. 1982. Correlations between holistic and quantitative methods of evaluating student writing, grades 4-12. ERIC ED 211 976.

Thorndike, R.L, ed. 1971. Educational Measurement. American Council on Education: Washington, D.C.

Thorndike, R.L. & E. Hagen. 1969. Measurement and Evaluation in Psychology and Education. Wiley: New York.

Thurstone, L.L. 1938. Primary Mental Abilities. Chicago University Press: Chicago, IL.

Tierney, R.J. & J. LaZansky. 1980. The rights and responsibilites of readers and writers: a contractual agreement. Center for the Study of Reading: Urbana, IL.

Tillman, M H. & L R. Veal. 1970. Systematic errors in rating and the quality of themes varying in mode of discourse and grade level. Mimeo.

Todd Trimble, M., L. Trimble & K. Drobnic. 1978. English for Specific Purposes: Science and Technology. English Language Institute, Oregon State University.

Toelken, B 1975. 'Folklore, worldview, and communication' in Ben-Amos & Goldstein, eds.

Torgerson, W.S. & B.F. Green. 1950. A factor analysis of English essay readers'. Research Bulletin 50-30. Educational Testing Service: Princeton, N.J.

Traxler, A.E. & H.A. Anderson. 1935. 'Reliability of an essay test in English'. School Review 63: 534-540.

Tuckman, B.W. 1972. Conducting Educational Research. Harcourt, Brace, Jovanovitch: New York.

Underhill, N, ed. 1982. 'The great reliability validity trade-off: problems in assessing the productive skills' in Heaton, ed.

Upshur, J.A. 1976. 'Discussion of a program for language testing

research' in Brown, ed.

Upshur, J.A. 1979. 'Functional proficiency theory and a research role for language tests' in Briere & Hinofotis, eds.

Upshur, J.A. & M. Fata, eds. 1968. Language Learning Special Issue No. 3: Problems in Foreign Language Testing. University of Michigan. .

Upshur, J.A. & T.J. Homburg. 1983. 'Some relations among language tests at successive ability levels' in Oller, eds.

Vachek, J. 1973. Written Language ; General problems and problems of English. Mouton: The Hague.

Valentine, C.M. 1932. The Reliability of Examinations. London University Press: London.

Valette, R. 1964. 'The use of dictee in the French language classroom'. Modern Language Journal 39: 431-434.

Valette, R. 1977. Modern Language Testing. Harcourt, Brace, Jovanovitch: New York.

Van Wagenen, M.J. 1920. 'The accuracy with which English themes may be graded with the use of English composition scales'. School and Society 11: 441-450.

Veal, L.R. 1966. 'Measuring writing improvement during an NDEA English Institute' Journal of Educational Measurement 3: 303-306.

Vernon, P.E. 1950 The Structure of Human Abilities. Methuen: London.

Vernon, P.E. & G.D. Millican. 1954. 'A further study of the reliability of English essays'. British Journal of Statistical Psychology 7: 131-142.

Vollmer, H. 1981a. 'Why are we interested in General Language Proficiency? in Alderson & Hughes, eds.

Vollmer, H. 1981b. 'Issue or non-issue: General Language Proficiency revisited' in Alderson & Hughes, eds.

Vollmer, H. & F. Sang. 1983. 'Competing hypotheses about second language ability: a plea for caution' in Oler, ed.

Vygotsky, L.S. 1962. Thought and Language. (translated by E. Haufmann & G. Vakar) MIT: Cambridge, Mass.

Wall, D. 1981. 'A pre-sessional academic writing course for postgraduate students in Economics'. Practical Papers in Language Education 4: 31-105.

Wallace, M. 1980. Study Skills in English. Cambridge University Press: London.

Wason, P.C. 1980. 'Specific thoughts on the writing process' in Gregg & Steinberg, eds.

Weaver, F. 1973. The Composing Processes of English Teacher Candidates: responding to freedom and constraint. PhD, University of Illinois at Champaign-Urbana, IL.

Weir, C. 1983. Identifying the Language Problems of Overseas Students in Tertiary Education in the United Kingdom. PhD, University of London.

Weir, C. 1986. 'Construct validity'. Paper presented at ELTSVAL Conference: British Council, London.

Weissberg, R. & S Buker. 1985. 'Strategies for teaching the rhetoric of Written English of Science and Technology' in Mackay, ed.

Wesdorp, H., B A, Bauer & A.C. Purves. 1982. 'Toward a conceptualization of the scoring of written composition'. Evaluation in Education 5: 299-315.

West, M. 1953. A General Service List of English Words with Semantic Frequencies and a Supplementary Word List from the Writing of Popular Science and Technology. Longman: London.

White, E. 1985. Teaching and Assessing Writing. Jossey-Bass: San Francisco.

White, E. & L.L. Thomas. 1981. 'Racial minorities and writing skills assessment in the California State universities and colleges'. College English 43: 276-283.

White, J. 1986. The Assessment of Writing. NFER-Nelson: Windsor, Berks.

Whitehead, F. The Disappearing Dais: a study of the principles and practice of English teaching. Chatto & Windus: London.

Whiteman, M.F. 1980. Writing: The Nature, Development and Teaching of Written Communication. Vol. 1: Variation in Writing. Lawrence Erlbaum Associates: Hillsdale, N.J.

Whorf, B.L. 1956. Language, Thought and Reality. MIT: Cambridge, Mass.

Widdowson, H.G. 1978. Teaching Language as Communication. Oxford University Press: Oxford.

Wilkinson, A.M. 1978. 'Criteria of language development'. Educational Review 30: 23-33.

Wilkinson, A.M. 1983. 'Assessing language development: the Crediton Project' in Freedman et al, eds.

Wilkinson, A.M. 1986a. The Quality of Writing. Open University Press: Milton Keynes.

Wilkinson, A.M. 1986b. The Writing of Writing. Open University Press: Milton Keynes.

Wilkinson, A.M., G. Barnsley, P. Hanna & M. Swan. 1980. Assessing Language Development. Oxford University Press: Oxford.

Wilkinson, A M. & M.E. Wilkinson. 1978. 'The development of language in the middle years'. English in Education 12: 42-52.

Williams, R. 1982. Panorama. Longman: London.

Williams, R. 1983/4. 'Using video to develop strategies in listening comprehension and examination answer writing'. English Language Research Journal 4: 50-67.

Williams, R.M J. Swales & J. Kirkman, eds. 1984. Common Ground: Shared Interests in ESP and Communication Studies. Pergamon Press: Oxford.

Willing, M.H. 1918. 'The measurement of written composition in grades 4 to 7'. English Journal 7.

Willing, M.H. 1926. 'Individual diagnosis in written composition'. Journal of Educational Research 13: 77-89.

Winkler, V. 1983. 'The role of models in technical and scientific writing' in Anderson et al, eds.

Winterowd, W.R. 1970. 'Topics and levels in the composing process'. College English 31: 828-835.

Winters, L. 1979. The Effects of Differing Response Criteria on the Assessment of Writing Competence. PhD, University of Califonia at Los Angeles.

Wiseman, S. 1949. 'The marking of English compositions in grammar school selection'. British Journal of Educational Psychology 19: 200-209.

Wiseman, S. & J. Wrigley. 1958. 'Essay reliability: the effect of choice of essay title'. Educational and Psychological Measurement 18: 129-138.

Witte, S.P. 1983a. 'Topical structure and revision: an exploratory study'. College Composition and Communication 34: 313-340.

Witte, S.P. 1983b. 'The reliability of mean T-unit length: some questions for research in written composition' in Freedman et al, eds.

Witte, S.P. & A.S. Davis. 1980. 'The stability of T-unit length: a preliminary investigation'. Research in the Teaching of English 14: 73-81.

Witte, S., P. Meyer, R. Cherry & M. Trachsel. Forthcoming. Holistic Evaluation: Issues, Theory, and Practice. Guilford Press: New York.

Witte, S.P. & R.E. Sodowsky. 1978. Syntactic maturity in the writing of college freshmen. ERIC ED 163 460.

Wood, R. & B. Quinn. 1976. 'Double impression marking of English language essays and summary questions'. Educational Review 28: 229-246.

Woods Chapman,C., L.J. Fyans & C. Thomas Kerins. 1984. 'Writing assessment in Illinois'. Educational Measurement: Issues and Practice 3: 24-26.

Wresch, W, ed. 1984. The Computer in Composition Instruction: A Writer's Tool. National Council of Teachers of English: Urbana,Illinois.

Wright, P. 1975. 'Presenting people with choices: the effect of format on the comprehension of examination rubrics'. Programmed Learning and Educational Technology 12: 109-114.

Wright, P. 1981a. '"The instructions clearly state..." Can't people read?' Applied Ergonomics 12: 131-141.

Wright, P. 1981b. 'Five skills technical writers need.' IEEE Transactions on Professional Communication 24: 10-16.

Yorio, C., K. Perkins & J. Schachter, eds. 1980. On TESOL 1979: The Learner in Focus. TESOL: Washington, D.C.

Yoruzuya, R. & J.W. Oller. 'Oral proficiency scales: construct validity and the halo effect'. Language Learning 30: 135-153.

Young, R E. 1978. 'Paradigms and problems: needed research in rhetorical invention' in Cooper & Odell, eds.

Young, R.E., A.L. Becker & K.L. Pike. Rhetoric: Discovery & Change. Harcourt, Brace & World: New York.

Zamel, V. 1976. 'Teaching composition in the ESL classroom: what we can learn from research in the teaching of English'. TESOL Quarterly 10: 67-76.

Zamel, V. 1982. 'Writing : the process of discovering meaning'. TESOL Quarterly 16: 195-209.

Zamel, V. 1983. 'The composing processes of advanced ESL students: six case studies'. TESOL Quarterly 17: 165-188.

Zamel, V. 1985. 'Responding to student writing'. TESOL Quarterly 19: 79-102.

APPENDIX A 1-3: THREE SETS OF WRITING TEST TASKS

A1: M2Q1
(texts follow)

Rubric for all M2 tasks (including both questions) was:

"Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Life Sciences:

Section 4 of your Source Booklet deals with the Green Revolution. Drawing on your own experience, discuss some of the advantages and disadvantanges of the introduction of modern farming techniques. (Write 15 to 20 lines.)

Medicine:

Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines.)

Physical Sciences:

New scientific processes often meet with opposition because of the pollution they cause to the environment, an example of which is referred to in Section 2 of your Source Booklet. As a scientist, how would you defend the continued use of such potentially harmful processes? (Write 15 to 20 lines.)

Social Studies:

Refer to the bibliography in Section 5.1. of your Source Booklet. If you had to read either Henry's *Who Lie in Gaol* or Size's *Prisons I have known*, state which one you would choose and write 15 to 20 lines giving the reasons for your choice.

Technology:

Look at the table on page 5 of your Source Booklet which describes certain characteristics of steel, iron and aluminium. Discuss the advantages and disadvantages of particular metals for a particular purpose with which you are familiar (e.g. building construction or vehicle manufacturing). Write 15 to 20 lines.)

General Academic:

same as M2Q1 SS (but Section 4 .I.)

APPENDIX A
M2Q1 : LS

Section 4: THE GREEN REVOLUTION

> And he gave it for his opinion, that whoever could make two ears of corn or two blades of grass to grow upon a spot of ground where only one grew before, would deserve better of mankind, and do more essential service to his country, than the whole race of politicians put together. JONATHAN SWIFT, *Voyage to Brobdingnag*, Part II, Chapter 7

Most readers will be familiar with the term 'Green Revolution' if not with the thing itself, for the public-relations job that has been done around this technology-package approach to UDC farming has been admirable. We will try to define it through a series of questions:

What does the term mean, technically speaking? It means breeding plants that will bear more edible grain—the 'two ears where only one grew before'—and thus increase yields without increasing cultivated crop areas. Traditional grains, especially those grown on the three poor continents, tend to be tall on the stalk for reasons of natural selection. That way they can get more sunlight, grow higher than the surrounding weeds, and resist flooding when heavy rains come. If one tried to produce double kernels on these long stalks, the plants would be top heavy, keel over and lodge in the soil. So the problem was to produce plants with short, tough stalks that could bear new fertiliser-sensitive hybrids. These dwarf varieties, capable of producing spectacular yields under ideal conditions, were eventually bred: they go under the name of high-yielding varieties, or HYVS for short. These plants can be adapted to any number of environments, but they are not as adapted as thousands of years of natural selection could make them—so they present problems of disease resistance. And they will not bear full fruit unless heavy doses of fertiliser are applied, and unless optimum irrigation is supplied. In other words, for us to get full benefit from the new 'miracle' seeds, they must have plenty of water, plenty of nourishment and plenty of chemical protection—pesticides and fungicides against disease; herbicides against the weeds that also thrive on fertiliser. The rub is that if a *single one* of these elements is lacking, HYVS can sometimes produce *less* grain than what could have been obtained with traditional varieties.

440

APPENDIX A
H201: ME

During the early days of my field research in the Edinburgh medical school, I was made aware of the fact that a number of students had previously worked in hospital settings—as nurses, orderlies, porters and so on. Thus I found myself conversing with students who were implicitly or explicitly comparing their ward work or bedside teaching with their own previous experiences of hospital life. Reports of such conversations soon found their way into my field notebook. For example, after one mid-morning coffee break, I noted the following:

'Arthur Gardiner and Harry Grant [pseudonyms] had both had some experience of mental hospitals. Gardiner said that he had once wanted to be a psychiatrist, but his experience had put him off. He had worked in a hospital in his home town, and the "old biddies" sitting around, looking up blankly (he imitated their vacant stare) had put him off psychiatry completely.

Harry Grant said that his experience with psychiatric patients had been happier. He recounted a story of a schizophrenic who started each day by declaring loudly the day of the week. "Tuesday morning", he would announce. "Mind you, that's just about all he did say", Harry added. He said he thought it was important to deal normally with psychiatric interviews: you can't start by asking, "Who's the King of England?"'

In themselves the remarks were pretty inconsequential, but their occurrence needs to be set against the background of their formal instruction. At the time when they were talking, the students were being introduced to taking psychiatric interviews, as part of their general introduction to clinical work. They had the task of taking such histories from patients in the general medical wards, and had small group sessions with a psychiatrist to discuss their 'findings' and also to explore their own reactions to this exercise. Such introductions to psychiatric work were a talking-point among many of the students. They debated among themselves a number of issues that arose. They discussed their own feelings on talking with patients on potentially distressing or embarrassing aspects of their private lives. Amongst other things, they questioned whether such activities were justified as 'purely academic' exercises. Several discussed their own unease at asking 'silly' questions in attempts to discover the patient's psychological status (like asking them if they knew what date it was, etc.). Again, there was disagreement over the validity of the psychiatrist's interpretations of the patients' replies, and indeed over the adequacy of psychiatric explanations in general. Some espoused a strong orientation towards organic explanations and tended to dismiss psychosomatic models as unfounded. Against this background of debate, then, the two students I was with over coffee set their own reactions within a context of previous personal experience. Thus Arthur Gardiner was dubious about the usefulness of the psychiatric work they were doing and the efficacy of psychiatry in

general. He partly justified his antipathy by reference to his past experience whilst working as a nurse in a psychiatric hospital. Similarly, Harry Grant was much more favourably disposed toward the specialty, and validated his attitude by reference to *his* experience. Whereas Gardiner picked on the depressing aspect of such work, Grant tended to emphasise what he saw as more endearing qualities of the patients' peculiarities.

As time went on, it became apparent that a large number of students had previously obtained some experience of work in hospitals, and were using this as a reference point in talking about their clinical instruction, and the problems they encountered in their work with doctors and patients. Thus, students came to discuss what they saw as problems in communicating with patients in the light of such previous experience. Again, this can be illustrated by an extract from my field notes.

'On the coach [from the hospital back to the medical school] I talked with Alan Pickering. I asked him what he was finding most difficult so far. He replied, "I don't want to say that the patients are stupid—but I find it very difficult to get through to them. I find it hard to pin them down." He explained that he found it difficult to phrase his questions to the patient in such a way as to get straightforward answers. People, he explained, were always rambling on about their own personal experiences.

He told me he had worked as a nurse previously, but the experience then had been totally different: as a nurse one encouraged the patients to talk at length about themselves. This, he thought, was a major function of the nursing role.'

Thus, the student's present difficulties were highlighted by reference to the hospital work he had already done. In particular, in this case we can note the implicit contrast between the work of the doctor and that of the nurse. Here it is exemplified by the student's perceptions of talk with patients. Having begun clinical medicine, as opposed to para-medical work, the purpose of his talk with patients is seen to differ. His communication with patients is now conceived in line with the doctor's position. What appears to have been learned from the nursing experience is not direct training for the clinical work of the fourth year—but rather some notion of the division of labour among hospital personnel. As I shall go on to describe, this is a major theme of students' prior exposure to hospital work.

At the end of the students' first year of clinical studies, I distributed a questionnaire concerned with their perceptions of the year's work (cf. Atkinson, 1973). As one item in that survey, I asked the students whether they had ever undertaken clinical work of some sort, as a nurse, porter or whatever. Additionally, I asked them if such experience had provided a useful grounding in interacting with clinicians and patients.

In all, 112 students returned completed questionnaires—just under 80% of the year group. Of those who replied, fifty-six—exactly half—had had a job of this sort at some time. Below I present some analysis of that item, and of the extended comments that students wrote on the general usefulness of such work.

In the first place, there was a sharp difference between the proportions

of male and female students who had undertaken such work (see Table 1).

TABLE 1 Proportions of male and female students who had had a 'clinical job'

	Male	Female
Had a 'clinical job'	37 (44%)	19 (68%)
No 'clinical job'	47 (56%)	9 (32%)
Total	84	28

The sex difference may arise from female students' easier access to temporary work in the strongly feminised area of nursing. Alternatively, it may reflect a sex difference similar to that described by Walton (1968) also for Edinburgh students. Walton describes the female students as tending to be more patient-centred than their male colleagues, who stress the more technical aspects of medical work. Thus the women may have sought out jobs that brought them into close personal contact with clinical work more frequently than the men. (In fact, both possibilities are reflections of culturally approved sex roles. The 'feminine' character of nursing and the female-associated patient-centred approach both depend upon traditionally stereotyped female characteristics of warm, nurturative interpersonal styles.)

Of the students who had taken such a job forty-eight (or 86%) believed that it had been of *some* value to them in understanding the clinical situation. But what appeared from students' comments was not that it provided directly applicable skills for doctor-patient interaction or the like. Rather, students tended to stress the insight that such work had given them into the general social functioning of the ward. They emphasised the knowledge that they had acquired of the routine ward work, which now provided the background for their activities. Also stressed was the degree of insight that had been gained into the position of patients in the hospital. In general, the attitude that emerged most strongly was that it had provided a view of clinical life from the other side—a sort of 'Upstairs, downstairs' perspective on hospital organisation and the work of its staff. From their fourth year on, the students will be primarily associated with the doctors teaching them. They will have crossed the divide that separates the 'medical' from the 'paramedical'.

APPENDIX A

M2Q1 : PS

Section 2: THE COMPOSITION OF THE AIR

Air is the mixture of gases which immediately surrounds the earth. It can be separated into its constituents by physical changes, such as liquefying the air by cooling and then allowing the temperature to rise. Each different gas will theoretically be vaporised from the liquid air at a different temperature. The actual industrial process is not quite so simple, since in order to obtain a particular gas with a high degree of purity several successive freezings and vaporisations are required.

The principal constituents of air are nitrogen, oxygen and argon, their proportions by volume being roughly in the ratio of 78:21:1. In addition there are very small traces of the inert gases helium, neon, krypton, radon and xenon. The proportions of the gases so far mentioned do not change greatly when different geographical locations are chosen for samples.

In addition to the gaseous elements previously quoted, air contains water vapour and about 0.03% of carbon dioxide. Air also contains impurities such as dust, soot and sulphur compounds, particularly near factories. Dry air has little effect on metals, but damp air, especially in the presence of sulphur compounds, such as those emitted by factory chimneys, has a severely corrosive effect on many metals.

442

APPENDIX A

M201:GA/SS

Section 4: A NOTE ON PRISON LITERATURE

THERE is a large literature on the subject of imprisonment. Memoirs by ex-prisoners are particularly common. The bibliography given here includes many works consulted in the preparation of the present book, but it is intended primarily as a guide for the general reader who wishes to study the causes and treatment of crime and prison conditions today and in the past. Readers without considerable experience in dealing with confirmed criminals would be well advised, when looking at any material written by ex-prisoners, to remember that the majority of such 'memoirs' are written heatedly and resentfully, usually including the most sensational incidents in the authors' experience, and often omitting any reference to positive, helpful treatment they received during their sentences. Moreover, prisons vary greatly in character, and the experience of one man in one or two prisons can never be taken as definitely typical of the treatment of all men and women serving imprisonment.

The author has been greatly helped, in compiling this list, by the staff of the Howard League for Penal Reform, which has an excellent library of penal literature, and by the Librarian of Kent County Library.

I. BIOGRAPHY

DENDRICKSON, G., & THOMAS, F.: *The Truth about Dartmoor,* Gollancz, 1952.

GREW, B.D.: *Prison Governor,* Jenkins, 1958.

The autobiography of a man with long and varied experience of prison administration.

HECKSTALL-SMITH, A.: *Eighteen Months,* Wingate, 1954.

HENRY, J.: *Who lie in Gaol,* Gollancz, 1952.

An ex-prisoner's account of her experiences in Holloway Prison, London, and at the open prison for women, Askham Grange, near York.

HIGNETT, N.: *Portrait in Grey,* Muller, 1956.

An account of prison life by a former coroner sentenced for fraudulent conversion. The author seriously under-estimates the idealism of members of the Prison Service, and his general picture of Wormwood Scrubs, where most of his imprisonment was spent, is distorted by bitterness. But it is an interesting companion to Mr Grew's book, which is largely concerned with the same institution at the same period.

HOWARD, D.L.: *John Howard: Prison Reformer,* Johnson, 1958.

An account of the eighteenth-century reformer's life and work.

SIZE, MARY: *Prisons I have Known,* Allen & Unwin, 1957.

A personal account of forty-seven years in the Prison Service, many of them as governor of prisons and Borstals for women and girls, with an excellent account of the opening of Askham Grange 'open' prison, of which Miss Size was first governor.

WHITNEY, JANET: *Elizabeth Fry,* Harrap, 1937.

An excellent biography of this remarkable pioneer.

WILDEBLOOD, P.: *Against the Law,* Weidenfeld & Nicolson, 1955.

A moving and sensitive account of the author's experience in Wormwood Scrubs Prison and of the incidents which preceded his conviction.

II. CRIMINOLOGICAL TEXTS

There are few English textbooks on Criminology directly related to our own penal system and our own social conditions. The work by Howard Jones listed below is the best brief introduction by an English academic criminologist. The others are American publications, and the very different social background of the United States and the peculiarities of its penal system should be borne in mind when they are used.

BARNES, H.E., & TEETERS, N.K.: *New Horizons in Criminology,* Prentice Hall, 1943.

JONES, HOWARD: *Crime and the Penal System,* University Tutorial Press, 1956.

RECKLESS, W.: *The Crime Problem,* Appleton-Century-Crofts, 1955.

SUTHERLAND, EDWIN H.: *Principles of Criminology,* Lippincott, 1934.

III. THE TREATMENT OF OFFENDERS

BENNEY, MARK: *Gaol Delivery,* Longmans, Green, 1948.

CALVERT, E. ROY: *The Lawbreaker,* Routledge, 1945.

EAST, DR NORWOOD, & HUBERT, W.H. DE B.: *The Psychological Treatment of Crime,* H.M.S.O., 1939.

ELKIN, W.A.: *The English Penal System,* Penguin, 1957.

A survey of the English Penal System in all its aspects, including a brief historical account.

FENTON, NORMAN: *The Prisoner's Family,* California: Atlantic Books, 1959.

FOX, SIR LIONEL: *The English Prison and Borstal Systems,* Routledge, 1952.

A classic account of the system and of official policy, by the present Chairman of the Prison Commissioners.

FRY, S. MARGERY: *Arms of the Law,* Gollancz, 1951.

GLOVER, ELIZABETH: *Probation and Re-education,* Routledge, 1939.

GRUNHUT, DR MAX: *Penal Reform,* Oxford University Press, 1948.

JONES, HOWARD: *Prison Reform Now,* Fabian Society, 1959.

KING, JOAN F.S. (editor): *The Probation Service,* Butterworth, 1958.

An account of the probation service by serving probation officers, including a description of the basic principles and methods used in case-work. The main aspects of a probation officer's duties (enquiries for courts, probation and supervision of offenders, after-care and matrimonial conciliation) are dealt with in some detail.

KLARE, HUGH J.: *Anatomy of Prison,* Hutchinson, 1960.

APPENDIX A

H2Q1 : TE

TABLE 2: SOME PROPERTIES OF METALS

material	*density* /kg m^{-3}	*proof stress 0·1 per cent* /N mm^{-2}	*tensile strength* /N mm^{-2}	*elongation on 55 mm* per cent	*modulus of elasticity E* /N mm^{-2}	*hardness (Brinell no.)*
stainless steels Fe : Cr : Ni : (Mo)	7850	210	510[i] 540[n]	50	207 000	170
high strength steel		350–430	495–617	19		150–180
mild (structural) steel			423–510	22		130
wrought iron			355	25–40		100
grey cast irons	7150	100–200	155–310	0·5–1·0	120 000	140–250
modular and malleable cast irons	7225	193–440	310–800	20–2·0	172 000	120–300
aluminium (Al) 99·0% pure	2650	—	70–140	2–20	68 300–72 400	23 (extrusions) 22–42 (sheet)
99·99% pure		—	80–100	3–45		15 (extrusions) 15–30 (sheet)

APPENDIX A 1-3: THREE SETS OF WRITING TEST TASKS

A2: SAPQ
(texts follow)

Rubric for all SAPQ tasks was:

"You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.
If you use the information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Life Sciences:

Explain how you would eliminate an undesirable genetic characteristic from a herd of cows. Refer to pp. 2/3 in your Source Booklet.

Medicine:

If you were a general practitioner in West Africa looking for information on how to reduce the mortality rate from bilharzia in your area, which of these books would you refer to and why? Refer to p. 9 in your Source Booklet.

Physical Sciences:

Discuss some of the ways in which air pollution can be reduced. Refer to p.3 in your Source Booklet.

Social Studies:

Describe the effects of the fall in death rates in Western Europe. Refer to pp. 3-5 in your Source Booklet.

Technology:

Name and describe three uses of bi-concave lenses. Refer to p.8 in your Source Booklet.

General Academic:

Explain why the 'green revolution' of high technology in food production has created serious social problems in India. Refer to pp. 7/8 in your Source Booklet.

APPENDIX A
SAPQ; LS

9446

How Good and Bad Characteristics are passed from Parents to Calves

Every cow and bull has in each body cell many pairs of genes. Each pair controls one particular character of the beast's make-up, horns, colour, etc.

One gene from each pair is carried by bull's sperm and cow's ovum at time of reproduction. These unite and thus the developing calf gets two genes for every character. Which gene each parent passes on is a matter of chance.

The two genes of a pair can be identical (thus, in the pair of genes controlling hair colour, they could both be for red) or a pair may be mixed (in which case one may be for red colour and one for white).

Where both parents have identical genes for a certain character it does not matter which of the pairs passes from parents to calf. The result will be the same. Their calves will always have identical genes for this character.

But if one parent has identical genes for a certain character and the other has mixed genes, then half the calves born will have identical genes for this factor and the other half will have mixed genes.

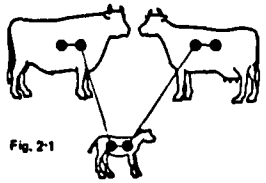


Fig. 2·1

Fig. 2·2

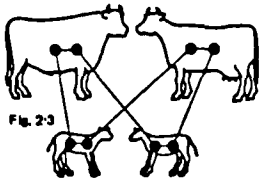


Fig. 2·3

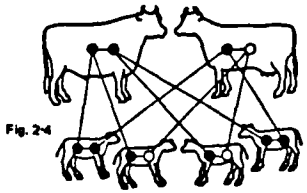


Fig. 2·4

Where both parents have mixed genes for a certain factor, then ½ of the calves will have identical genes of one kind, ¼ will have identical genes of another kind and ¼ of the calves will have mixed genes.

So far, the breeding examples quoted have been on simple body characters like hair colour controlled by one pair of genes. Milk production is determined not by one pair of genes but by many pairs each controlling various factors. These genes cannot be examined to determine whether they are identical or mixed.

The only way of finding out whether animals breed true for milk is by knowing whether all their progeny consistently produce good yields. This is a sound indication that the parents have identical genes for many of the factors influencing milk production.

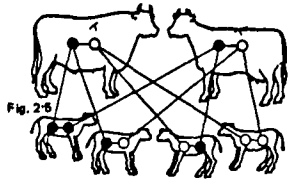


Fig. 2·5

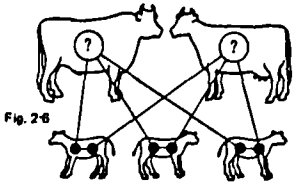


Fig. 2·6

APPENDIX A
SAPQ : ME

Section 6: BIBLIOGRAPHY

Belcher, D.W. and others
A household morbidity survey in rural Africa.
International Journal of Epidemiology, 1976, **5**, 113–20

Browne, S.G.
Research in a 'bush hospital' in Africa.
Tropical Doctor, 1976, **6**, 187–89

Burton, J.H.
Problems of child health in a Peruvian shanty town.
Tropical Doctor, 1976, **6**, 81–83

Chandrasekhar, U., Nandini, S. and Devadas, R.P.
Protein quality and acceptability of CARE's Kerala Indigenous Food.
Indian Journal of Nutrition and Dietetics, 1976, **13**, 1–6

Conacher, D.G.
Medical care in Ethiopia.
Transactions of the Royal Society of Tropical Medicine and Hygiene, 1976, **70**, 141–44

Davies, J.C.A.
The organisation of tuberculosis service in the Midlands of Rhodesia (1963–1972).
Central African Journal of Medicine, 1976, **22**, 74–78.

Ebie, J.C.
Towards improving the administrative machinery for health care in the Mid-western State of Nigeria.
Nigerian Medical Journal, 1976, **6**, 112–17

Giel, R. and Harding, T.W.
Psychiatric priorities in developing countries.
British Journal of Psychiatry, 1976, **128**, 513–22

Gunaratne, V.T.H.
The challenge faced by the medical profession in tropical developing countries.
Tropical Doctor, 1976, **6**, 180–84

Levi, G.
Health. An integral part of development [*in developing countries*].
Nursing Mirror, 1976, **143**, No. 6, 63–66

Lowry, M.F., Howell, V. and Bird, S.
Paramedical assessment of gestational age in the newborn.
West Indian Medical Journal, 1976, **25**, 17–22

Lucas, A.O.
Surveillance of communicable diseases in tropical Africa.
International Journal of Epidemiology, 1976, **5**, 39–43

McDowell, J.
In defence of African foods and food practices.
Tropical Doctor, 1976, **6**, 37–42

Section 2: THE COMPOSITION OF THE AIR

Air is the mixture of gases which immediately surrounds the earth. It can be separated into its constituents by physical changes, such as liquefying the air by cooling and then allowing the temperature to rise. Each different gas will theoretically be vaporised from the liquid air at a different temperature. The actual industrial process is not quite so simple, since in order to obtain a particular gas with a high degree of purity several successive freezings and vaporisations are required.

The principal constituents of air are nitrogen, oxygen and argon, their proportions by volume being roughly in the ratio of 78:21:1. In addition there are very small traces of the inert gases helium, neon, krypton, radon and xenon. The proportions of the gases so far mentioned do not change greatly when different geographical locations are chosen for samples.

In addition to the gaseous elements previously quoted, air contains water vapour and about 0.03% of carbon dioxide. Air also contains impurities such as dust, soot and sulphur compounds, particularly near factories. Dry air has little effect on metals, but damp air, especially in the presence of sulphur compounds, such as those emitted by factory chimneys, has a severely corrosive effect on many metals.

APPENDIX A
SAPQ : SS

Section 2: HUMAN POPULATION

Para. 1 It is probably well realised now that the very great population increases during this century, and particularly since the close of the Second World War, are *not* the result of an increase in human fertility, but rather of a decline in *mortality* resulting from advances in, and the wider application of, *modern medicine*. It is striking to realise that whereas it has taken the world 200000 years to attain a population of 2500 million, it will now only require thirty years to add a further 2000 million.

Para. 2 It seems that the modern phase of accelerating population increase began during the seventeenth century and was well under way in the eighteenth century. The sharp upward turn in the rate of population increase during this period may be related to the striking advances made in the fields of agriculture, industry, medicine and sanitation. In these the countries of Western Europe were in the forefront. Between 1650 and 1900 Europe's population, despite considerable emigration, multiplied itself four times and its share of the world's population increased from 22 to 27 per cent. Asia's population grew at a slower rate and by 1900 the increase was about three times. By the end of the first half of this century Europe's population had increased almost six-fold since 1650 and Asia's population had quintupled. Thus the rate of Asia's population increase has gone up appreciably during this century. Admittedly, higher rates of increase since the seventeenth century have been recorded in the relatively empty lands of the Americas, but the numbers involved have been relatively small; four-fifths of world population is now in Europe and Asia.

Para. 3 The period of really critical increase in the rate of population growth has been the last three decades. Until 1940 the world's annual increase of population was 1·0 to 1·2 per cent. This quickening in the rate of growth has not yet been checked; the situation will worsen before we can hope for easement. This recent excessive acceleration has been due to a series of scientific and medical advances whose application has resulted in what has been termed 'death control'. The vigorous introduction of medical services, new drugs, instruction in hygiene and improved sanitation into the poorer countries has often markedly extended the expectation of life. In Britain the expectation of life at birth is about 70 years, in India at the beginning of the century it was only 24, but by 1960 it had increased to 41 years (Table 1). Postponement of death has been particularly successful in the case of infants, and infant mortality rates in the under-developed countries have started to show astonishing reductions. During the period 1948–67 Sri Lanka's mortality rate fell from 92 to 48 and Chile's from 147 to 92 deaths under the age of 1 year per thousand live births—examples typical of most of the under-developed countries.

Table 1: *Expectation of life at birth in selected countries*

	(years)				
	1900	*1930*	*1950*	*1960*	*1970*
United Kingdom	52	—	69	71	71
France	47	—	65	70	72
Bulgaria	40	46	—	66	71
Japan	44	—	58	70	72
India	24	27	32	41	—
Chile	—	37·5	52	57	—
Mexico	—	33	—	57	62
Kenya	—	—	—	—	44
Egypt	—	—	—	53	—

The average of male and female life expectations is given.
SOURCES: United Nations, *Demographic Yearbook, 1953*, Table 19; *1958*, Table 31; *1964*, Table 23; *1971*, Table 3.

Para. 4 As a result of social and economic developments populations usually pass through a number of distinct stages in their growth. In the *pre-scientific* period both birth and death rates were high: population increase was slow and irregular. This period is known as the 'high fluctuating' stage and in this country was coming to an end early in the *eighteenth century*. As medical knowledge and sanitation improved, the next phase occurred: the 'early expanding' phase of rapid increase. In this stage death rates fell markedly but birth rates remain both high and constant and a maximum increase of population occurs. The third or 'late expanding' stage finds death rates continuing their fall, but, responsive to rising standards of living, families become smaller and a sharp decline in the birth rate sets in: population still increases, but at a less rapid rate. Finally in the 'low fluctuating' phase, both birth and death rates steady at a low level (in Britain at about 16 and 12 per thousand respectively), increase still takes place, but very slowly and the population seems to be reaching a phase of stabilisation. These conditions for England and Wales are demonstrated in Fig. 3.

Para. 5 It may well be thought that the developing lands currently receiving the benefits of advanced agriculture, industry and medicine are now passing through the early stages of the population cycle and that a pattern similar to that of Western Europe will ensue. However, the problems of these countries, and indeed the world's population problem as a whole, arise from the very different timing and telescoping of the various phases.

Para. 6 First, it must be realised that in Western Europe, the type area, change was gradual. The knowledge and application of death control arose from a long period of trial and error. The slow but

APPENDIX A
SAPQ : SS

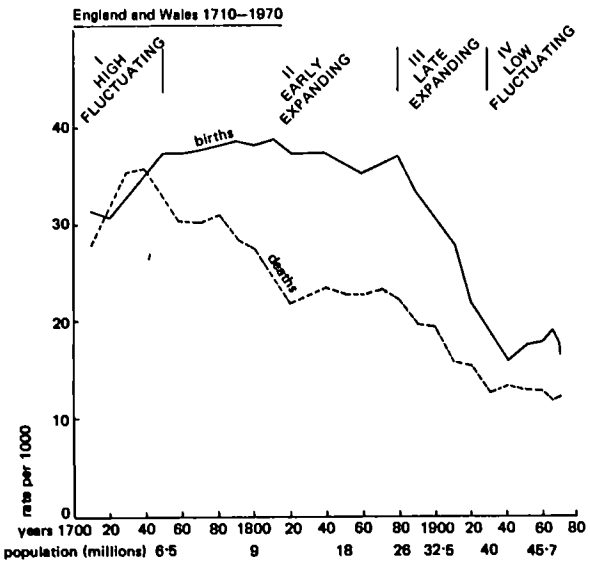


Fig. 3. The population cycle.

steady accretion of population both stimulated and allowed full advantage to be taken of the new developments in industry and transport: in short, they permitted economic expansion which throughout the nineteenth century raised standards of living, making possible, at a price, the choice of comforts and amenities and thus fostering a material desire to limit families. In England and Wales this had started to take effect in the decade 1871 80, when the birth rate, dropping from 35·5 to 34·1, started a downward plunge that was not arrested until during the Second World War. The situation is very different in the developing countries where Europe's hard-won knowledge of two centuries is readily, and relatively instantaneously, available. In the reduction of death rates, what has taken Western Europe two centuries to accomplish is now occurring in one or two decades in parts of Asia and Latin America. This is graphically illustrated by the much-quoted example of Sri Lanka, where DDT and a well-organised campaign wiped out the malarial mosquito in seven years (1945 52) and the death rate fell from 22 to 10 per thousand. It took about 300 years to clear malaria from England and the comparable fall in the death rate was spread over seventy years. During this time, however, England's birth rate fell from 35 to 15 per thousand

whereas that of Sri Lanka remains above 30 and the population is increasing at a rate that will double it in thirty years. Here is the explanation of the population explosion now becoming manifest: birth rates still in stage 2 of the population cycle, out of phase with falling death rates often already approaching stage 4.

Para. 7

These are fundamental differences from the population cycle of Western Europe and it is possible that a new cycle applicable to the developing lands will be formulated. These differences are particularly manifest in the structure of these populations and in the character of the economic development they necessarily engender. Development, particularly industrial development, will not be an evolution from within (as in Western Europe) but will be an imposition from without. The scale of development must be very different since the initial populations are more vast and the standards of living lower than those of Western Europe two centuries ago and the greater these populations the greater the rate of growth of total output needed to raise the standard of living.

450

APPENDIX A

M201 : TE

Section 4: LENSES

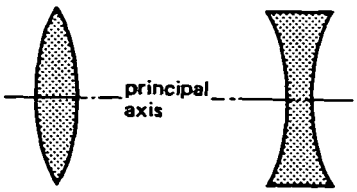
The refraction of light is utilised in a variety of ways that may be of considerable scientific benefit. A large proportion of these ways involve light passing through a lens or a series of lenses. The lens gets its name from the Latin word for a bean, because the shapes of the commonest lenses are similar to those of beans or lentils.

A lens is a piece of glass or other transparent material whose thickness varies from the middle to the edges, bounded by curved surfaces on one or both sides.

Since very early times, lenses have been used to bring together rays of light in a concentrated form. They were originally known as 'burning glasses' because the sun's rays could be concentrated to such an extent that sufficient heat could be generated to start a fire.

Lenses are used in spectacles to improve vision, in microscopes to make very small objects easily visible, in telescopes to make distant objects appear near, and in cameras and projectors to produce a sharp image on a film or a screen.

There are a large number of different shaped lenses in common use, but for convenience they may be grouped under two headings—converging or diverging lenses. *Converging lenses* cause rays of light to come together after passing through them, and *diverging lenses* cause rays of light to spread out after passing through them.



(a) (b)

Fig. H5 (a) Bi-convex lens (b) Bi-concave lens.

Figure H5 shows examples of each of these types of lens. (a) is what is known as a bi-convex lens, because both of its surfaces curve outwards. The surfaces can have the same or different radii of curvature, depending on the use of the lens. (b) is known as a bi-concave lens, having both of its surfaces curving inwards. Again the radii of curvature can be the same or different in a bi-concave lens. (a) is an example of a converging lens and (b) is an example of a diverging lens. In such lenses the line passing through the centres of curvature of the lens surfaces is known as the *principal axis* of the lens.

APPENDIX A

M2Q1 : GA/SS

Section 4: THE VIOLENT HARVEST

To loud acclaim from the prestigious international audience, Dr. Norman Borlaug advanced to the podium to receive the Nobel Peace Prize. It was 1970 and the prize was a generous gesture towards an agronomist who, pottering about in various scientific greenhouses, had bred new and fabulously prolific varieties of wheat and rice. Deployed in India over the previous five years the 'miracle' seeds had helped produce record harvests. The Nobel judges made the understandable connection that the creation of more food in the subcontinent went hand in glove with peace.

But it is violence not peace that is being harvested in India's fields.

The new varieties of seeds that have been so profusely scattered are rather like highly bred dogs. They need to be pampered or they sicken and die. The new seeds have to have regular supplies of water — so only irrigated fields can be planted. They have to be bedded down with expensive artificial fertilisers. The intensive hothouse breeding means that they are vulnerable to disease and need to be cared for with pesticides. And finally all the cosseting is only worthwhile on a large scale and generally with the help of machines.

The idea of the green revolution was embraced enthusiastically by the New Delhi government. Their backs were against the wall. The number of hungry Indians was increasing remorselessly. And although progressive taxation would have helped provide food for nearly everyone, that was political dynamite.

It was a far softer option to seize on new and scientific ways to increase the size of the cake rather than enforce fairer slices.

The 1965 Indian Five Year Plan swung a lot of government money behind the new seeds. All the elements of the technological package were provided *without* tampering with the basic pattern of 'who owned what' in the countryside. It was new wine into old bottles.

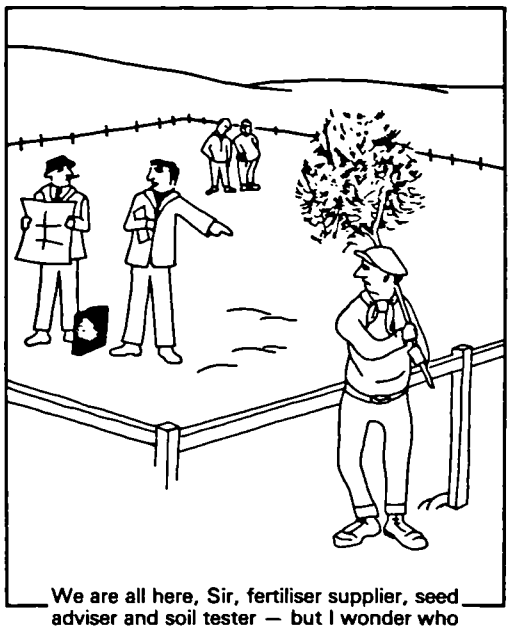
Scarcely ever in agricultural history has there been such a rapid transplant of new farming technology, on such a massive scale and with so great a success. By the end of the decade the number of tractors being used had increased five-fold, tubewells for irrigation 38 times, the area sown with the new seed from two million to twenty-two million hectares.

The green revolution in India reached its highwater mark in 1970/71 when a record-beating crop of over 100 million tons of foodgrain were harvested.

But the achievements have turned sour.

A farmer has to make a sizeable investment and take a risk for the new seed to pay off. The government supplied cheap credit for tractors, expensive seed and fertilisers to be bought. But it was the well-off farmers who were good risks and qualified for the loans. It was the well-off farmers who understood the complicated paraphernalia which surrounded the planting of the new seeds. It was the well-off farmer who had more than enough land to provide for his family's food needs and could risk giving the new seeds a whirl. It was the well-off farmer who had the irrigated land. And it was the well-off farmer who had large enough fields to make a tractor a worthwhile asset.

Tenant farmers were squeezed off the irrigated land. It now became more profitable to farm with the new technology than collect half the small-holders' crops in rent. Small farmers with their regular crops found that market prices had been driven down by the bumper harvests of the large



We are all here, Sir, fertiliser supplier, seed adviser and soil tester — but I wonder who that man is over there.

452

landowners. And the green revolution beneficiaries' extra cash was ploughed back into buying up plots from the small fry sunk into debt.

The gulf between the village rich and poor has widened by leaps and bounds. There might have been more food on the market. However, many people have less money to buy it.

But to engage in the primitive Luddism of technology-bashing is daft. The agricultural work of Norman Borlaug is a breakthrough But it can only become a force for peace once the agricultural social structure has been changed

A3:GAPQ

Read the text below and then answer the question under it.
You have 35 minutes altogether for reading and writing. Use the back of the Q1. paper for notes if you wish.

British still believe in sin, hell and the devil

by JUDITH JUDD

MOST Britons still believe in the concept of sin and nearly a third believe in hell and the devil, according to the biggest survey of public opinion ever carried out in the West.

Britons have a stricter moral code than their fellow Europeans, especially about sex under the age of consent, fiddling the dole and keeping money they have found. But they are more permissive about euthanasia and failing to report accidental damage to a parked vehicle.

The findings of the survey, begun in 1978 in nine western countries, show that belief in sin is highest in Northern Ireland (91 per cent) and lowest in Denmark (29 per cent). More than twice as many Americans as Europeans believe in hell and the devil.

Even 15 per cent of atheists believe in sin and 4 per cent in the devil.

A preliminary analysis of the findings, to be published in a book in the autumn, is given in the Roman Catholic weekly, *The Tablet*. It shows that 78 per cent of Europeans think there is good and evil in everyone.

The Irish have the most optimistic view of mankind. They think 34 per cent of people are basically good. The figure for the French, who take the most jaundiced view, is 5 per cent.

Most Europeans admit that they sometimes regret having done something they felt was wrong. The Italians and Danes suffer most from such regrets, the French and the Belgians least. The rich regret more than the poor.

The survey, which was carried out by an international team of academics, examines the 'sins' recognised in the West. The Ten Commandments, apart from those about Sunday and worship, are still rated highly.

Killing is top, followed by stealing and honouring parents. Britons rated the prohibitions on adultery and coveting thy neighbour's wife higher than did any other nation.

Most of those questioned cited honesty as the most important quality to be encouraged in children. Only the British put good manners second. For other nations, tolerance and respect for other people came next.

The rich are less likely to believe in sin than the poor. The right takes a more cheerful view of the nature of man than the left.

Among parents the strictest attitudes were found among believers in God and regular worshippers. Left-wing parents are less strict than right-wing parents. Parents in lower income groups are tougher disciplinarians than their wealthier counterparts.

Professor Jan Kerkhofs, a Jesuit priest at Louvain University, in Belgium, who is director of the project, said last week that between 1,200 an 2,000 people had been questioned in each country and the findings were still being analysed.

(The Observer: 27/2/83)

Which of the 'sins' mentioned in the text do you think are most serious, and why?

APPENDIX B: ORIGINAL SCORING PROCEDURE

Marking the Test

Marking of Multiple-Choice Sub-Tests G1, G2, M1 (all modules/patterns, except for the Non-Academic Training Module)

Cross through any multiple responses. These occur when a candidate has underlined more than one letter for any one item. They are always counted as wrong, even when one of the letters is correct. It may be clear that the candidate has changed his/her mind but not completely erased his/her first choice. In these cases the marker should use his/her discretion.

Check that the correct template has been selected according to the Version used, and for M1, remember that the template must be selected according to the module used. Place it on top of the Answer Sheet, taking care to align them accurately. Taking one column at a time, count the number of underlined letters in the template boxes.

Enter the figure against 'Raw Score' in the section for office use. Then, using the table printed on the template, convert this figure into a band and enter it against 'Band'.

The scoring should be double-checked, if possible by someone else. The section headed 'Comments' should only be used to indicate something which may have significantly affected a candidate's performance, e.g. 'arrived late'.

Marking of M2 (all modules/patterns, except for the Non-Academic Training Module).

Question 1

Assess with the aid of the M2 Topics ~~and Model Answers~~ and Q1, Writing Assessment Scale and Writing Samples provided (see Appendix C pp 13-22). Use only whole, not half, bands. Judge according to the communicative quality of the writing, the effectiveness with which the arguments are presented, the logical structure of the presentation and the accuracy and appropriateness of the language used. Candidates should not be heavily penalised for making factual errors in a subject with which they may not be familiar, but answers should be relevant to the questions asked.

Remember that it may not be possible or sensible to expound a specialist topic wholly in one's own words. Question 1, however, has been worded so that it should not encourage answering by wholesale lifting from the text. Wholesale lifting should be assessed as band 1 (see M2 Writing Assessment Scale). Partial lifting may contribute to an appropriate answer and should be assessed accordingly.

Question 2

The accuracy of the information given is more important than in Question 1 and for this reason Model Answers are given on pages 13-18. The communicative value of the writing and the correctness of the English used should also be considered. The purpose of the Model Answers is to guide those assessors who may not be familiar with the subject matter. They should not be interpreted as definitive models of the presentation: they offer a useful, though by no means exhaustive, summary of the data that may be expected in an acceptable handling of the topic.

BAND	BRIEF PERFORMANCE DESCRIPTION
9	**Expert Writer:** theme presented in a readable, intelligible, logical and interesting manner. Writes with complete accuracy and in the appropriate style. The reader is given a sense of mastery of the language and of the ability to handle the topic with complete competence.
8	**Very Good Writer:** theme presented clearly and logically, with accurate language forms and good style. Only very occasional inaccuracy or inappropriacy but which does not affect the communication. The reader can follow with no strain and will appreciate the argument expressed.
7	**Good Writer:** theme presented in a well-ordered, intelligible manner with well-structured and relevant supporting detail. Generally accurate in language and appropriate in style, but occasional lapses can affect the communication on first reading. The reader has, however, the impression of a functionally efficient writer.
6	**Competent Writer:** theme presented fairly logically and intelligibly. Reasonably accurate use of the language system. May have inaccuracies of style and presentation but showing an adequate functional competence. Can be read with only occasional strain put on comprehension.
5	**Modest Writer:** theme can be followed, but logical presentation may be broken and lack clarity or consistency. Several inaccuracies and style not always appropriate to presentation. May lack interest or variety, but the basic message is presented. The reader will have to strain on occasion to comprehend meaning.
4	**Marginal Writer:** theme can be followed with effort, and closer reading reveals lack of logical structure, clarity and consistency. Inaccurate vocabulary and sentence use coupled with inadequate connectors and cohesive features. Elements of information required may be omitted, repeated or inappropriately expressed. The reader has general difficulty in working out the message, though can eventually do so.
3	**Extremely Limited Writer:** elements of the information required are provided, but the presentation lacks any coherence. Uses over-simple sentence structure and impoverished vocabulary with continual errors and inappropriateness. Below level of functional competence though the reader may work out the general message.
2	**Intermittent Writer:** elements of the information required not provided, although a general meaning comes through intermittently. Either copies or produces strings of words. No real communication with the reader having constant problems in making out any message.
1	**Non-Writer:** cannot write the language. OR: cannot be adequately assessed either because answers have been lifted 'en bloc' from the Source Booklet, or because a clearly irrelevant stock answer has been produced.
0	No questions have been attempted.

APPENDIX B

Band 8: Very Good Writer

The dentition of each mammal is determined by its feeding habits and diet. The sheep differs from the dog in its dental formula, in particular, the absence of canines and ~~the~~ upper-jaw ~~canine~~ incisors, and in the manner of jaw movement. The premolars and molars serve a grinding function with the lower jar taking a circular path about the jaw point. In the dog, the emphasis shifts to the canine and carnassial teeth which are used to ~~attack~~ chop up meat into chunks, ~~and are quickly~~ Movement of the lower jaw is in the vertical plane about the jaw point in the dog.

Band 7: Good Writer

Though interior heating is not necessary for a large part of the year in a Tropical country such as India, heat pumps can still be gainfully employed not only during the colder periods of the year but also in those instances where heating is necessary, irrespective of the weather. With more efficient and economically viable heat pumps, the refrigerator, hitherto regarded as a luxury item, can be brought to the lower income homes.

In places like Delhi, Srinagar, Jaipur and other such places in the North, where the weather is quite cold in the winter, heat pumps can be used to warm up the rooms or buildings

457

Band 6: Competent Writer

As a plant breeder, I will join the rest people to produce high yielding, well adapted and many disease resistant varieties for my Country. This is indeed a great benefit which is very likely to put my Country in a very good stead as one of the leading agricultural countries in the world. Without my coming to Britain, the acquisition of the technology used in producing hybrids, which has been a giant step in crop improvement, could have probably not come to me. This is not because the course is not done in other countries but language has always been the problem. In Britain I have been introduced to modern techniques and I have shared knowledge with a well informed supervisor. All these will be of considerable help to me in future in my endeavours to improve agriculture in my Country.

Band 5: Modest Writer

I ~~like~~ shall choose to read sizes 'Prisons I Have Known". Because here I may get the real information of the Prison source. The author described here own experience and it is about the female prisoners. She was the Governor of the 'Open' prison. It is also interesting thing that the prison is open. On the other hand 'Portrait in Grey' by Hignett is not well written He under-estimated the idealism of members of the prison source. I do not, perhaps get any real picture as information in it.

Band 4: Marginal Writer

first of all a Iam very interested to lear
ang English language. I believe that a good langu
age could be learn in own country with mixing
British people, seeing their own caulcha and
customes. I always had in diffuculty to read
and understand ~~as~~ medical books in English.
As we know a lot of medical books are
written in English. A lot of medical researcs
are done in ~~Book~~. England and U.S.A. And
of course their puplished in English. In my
own country if you don't know any
~~as~~ foreign language as a doctor you have
to wait somebody translate a book from English
to Turkish. This may take years.

Band 3: Extremely Limited Writer

If pt with Cancer in The breast and had been
removed and She develp any sing of depres. first
of all to tell her the True about The tumb which
removed and give her a hope of complete quiet
life. If it benin and if it is malegnant, give her a
proper treatment and not to loss a hope for the best

Band 2: Intermittent Writer

she must go to Doctor who is speacliced
and to detail of the oprostion is nessery to do
it it depend on Mr doctor who will going to
treat her.

Appendix C

Selection of Modular Options

Where some indication is needed as to which modular area a prospective candidate should choose, the checklist below is offered as a rough, and by no means exhaustive, guide to the more obvious professional fields that may be assumed to fall under the various disciplinary headings of the test:

Life Science Module

agriculture
animal husbandry
veterinary studies
rural science
biology
botany
ecology
zoology
fisheries
forestry
genetics
food sciences

Non-Academic

auto/motor engineering
electrical engineering
ambulance service
fire service trainees
plumbing
carpentry
sewage engineering

Medicine Module

medicine/surgery
dental surgery
basic medical sciences
psychiatry
pharmacology

Physical Sciences Module

physics
chemistry
astronomy
mathematics
crystallography
operational research
optics
nuclear science
computing

Social Studies Module

social sciences/services
management - all types
administration (social/public)
government (central/local)
planning - all types
industrial relations
political science
population studies

Technology Module

engineering - all types
technology - all types
building
surveying
hydrology
metallurgy
soil mechanics

General or interdisciplinary cases, or those not easily accommodated under any of the above mentioned headings, should be assigned to the General Academic Module.

Development of the Tests

APPENDIX D

Multiple-choice sub-tests

The general tests (G1 and G2) and the first test of each module (M1) are in multiple-choice form. They were constructed with reference to the discussion document **Specifications for an English Language Testing Service**, published by the British Council in 1978. Materials were prepared by small teams of writers and subjected to a pilot test on approximately 950 students in 34 centres in the United Kingdom. They were subsequently edited and pretested on a total of 603 students in 19 countries overseas before being cast into their final form for administration to candidates in 1980. Table 1 gives the statistical parameters of the tests, calculated for approximately the first thousand candidates for the Service.

Table 1

Test	Number of Candidates	Number of Items	Options per Item	Mean Score (%)	Standard Deviation (%)	Reliability[1]
G1 - Reading	1087	40	4	55	22	.90
G2 - Listening	1073	35	4	52	21	.88
M1 - General Academic	172	40	4	41	19	.87
M1 - Life Sciences	157	40	4	49	19	.84
M1 - Medicine	131	40	4	55	21	.87
M1 - Social Studies	227	40	4	44	20	.88
M1 - Technology	200	40	4	64	23	.92
M1 - Physical Sciences[2]	-	-	-	-	-	-

Score Reporting

Performance in each test is reported in the form of a band, the significance of which is contained in published band descriptions. These descriptions are applied directly to the candidate's answers in the case of the subjectively marked tests M2 (Writing) and M3 (Interview). Band cut-off points for the multiple-choice tests were determined from the pretest results by matching the band distribution for each test with the M2 band distribution for the same candidates. In this way, the average candidate can be expected to have a flat profile in the sense that his band is the same on G1, G2, M1 and M2. Candidate profiles will of course depart from this pattern, but the departure will reflect variation in performance across the tests when compared with the average, and not differences in the intrinsic difficulties of the tests. Table 2 shows percentages in the sample above falling into the various bands.

Table 2

Test	BAND						
	0-2.5	3-3.5	4-4.5	5-5.5	6-6.5	7-7.5	8-9
	%	%	%	%	%	%	%
G1 - Reading	9	18	29	27	14	1	2
G2 - Listening	7	15	21	25	16	12	4
M1 - General Academic	4	19	33	20	16	6	2
M1 - Life Sciences	9	29	25	21	11	6	-
M1 - Medicine	14	17	29	30	7	2	1
M1 - Social Studies	15	22	27	15	15	4	1
M1 - Technology	8	10	15	18	20	13	14
M1 - Physical Sciences[2]	-	-	-	-	-	-	-

1. The estimate of reliability is derived using Formula 20 of Kuder and Richardson.
2. Insufficient numbers.

1985 ASSESSMENT GUIDE : APPENDIX E

ENGLISH LANGUAGE TESTING SERVICE

CRITERION SCRIPT 2

Band

M2 (WRITING) TEST · MEDICINE MODULE

Test Centre ...EDINBURGH U..... Date .11 JAN 1984.....

Candidate's Name .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines).

The advantages and disadvantages that a medical student or a doctor could have had by having some previous experience of hospital life depends upon their respective duties and the skills they had to employ. Thus, those who come more frequently in contact with patients have a better chance of establishing rapport with them provided the doctors do not feel themselves standing on a pedestal. Secondly, doctors with previous paramedical experience are well aware of the hospital organisation and those previously practising technical skills adapt nicely to situations which are a continuation of their previous skills. Moreover, they now understand and are able to explain many diseases and procedures, which are nothing new to them.

I think that the only disadvantage for a previous hospital experience as paramedical personnel is not getting used to the idea that one has to work with a broader perspective, maintaining the humility which one has cultivated previously by occupying a less rewarding position of a paramedic.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

462

TURN OVER FOR QUESTION 2

ENGLISH LANGUAGE TESTING SERVICE

CRITERION SCRIPT 6

Band

M2 (WRITING) TEST GENERAL ACADEMIC MODULE

Test Centre EDIN. Date 5/10/83

Candidate's Name ...

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 5.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have known*, state which one you would choose and write 15 to 20 lines giving the reasons for your choice.

If I had to read either 'Henry's 'Who lie in Gaol' or Size's 'Prisons I have known', I would prefer to choose the second one. 'prisons I have known'. Because:

1. It is a personal account. personal account are more reality, more attractive.

2. It has been quet a long time for she wrote the book Forty-seven years for a worker, there must be a lot of experience. so the book must be in good detail concerning the prisons.

3 She was the first governor of open prison. She must be very intelenge, and did very work in her position

4. she had been in many prisons as a governor. she must have many knowledge about the prisons and how to rule the prison.

In addition to that. she have an excellent account of the open prison. so I would like to choose this book

463

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

APPENDIX E

GLISH LANGUAGE TESTING SERVICE

Band

CRITERION SCRIPT 9

M2 (WRITING) TEST LIFE SCIENCES MODULE

st Centre ..... University of Edinburgh ..... Date ..... 21 September 1983

ndidate's Name ....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

estion 1: Section 4 of your Source booklet deals with the Green Revolution. Drawing on your own experience, discuss some of the advantages and disadvantages of the introduction of modern farming techniques. (Write 15 to 20 lines).

Now, many people in the world who live in Asia and Africa lack of their foods. So, many people are dying because of the shorts of their foods. So, I think it is important to increase the mass of cultivated crop. The modern farming is the good way to resolve above the problem.

But I think it is difficult to make a new seeds which satisfy our purpose. If we can get it, perhaps it won't be perfect.

Now, people in progresive country have luxialy meal. if they do half to meal. above problem is a little resolved.

NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

464

RN OVER FOR QUESTION 2

APPENDIX F: TRANSCRIPT OF RATERS' DISCUSSIONS

It should be noted that:

1) because these data were originally collected for another purpose, i.e., the development of the scoring guide, not every part of every discussion was recorded: some discussions will therefore seem incomplete
2) the recordings have been transcribed with many of the hesitations and false starts, but where these did not appear to have any semantic significance some of them have been condensed to make the transcripts less broken up and easier to read

Writer Number 3

Rater B: I started with 6-7-8...there are certain ....hum....areas in it where ... he hasn't used anaphora in the right way, for example or he, or she, ... hasn't included ... 2 parts of the sentence there should've been an 'also' to make it balance properly ... maybe the slip-up on 'reason' plural- this kind of thing ... just a ... a number of little things, though otherwise ... it reads very well ... I wondered too it crossed my mind ... that perhaps it's easier ... for these people who are lifting ... um ... quite ... impressive-sounding terms directly from the text, you know ... coefficients of various things ... um... so I tended to ... then go for a 7 on that ... 'cos it is reasonably well-organised and it does directly answer the question

Rater C: I found it difficult to assess for the reason that I didn't fully understand the content ... I feel it is ... a lot easier to write this sort of ... sophisticated jargon ...

Rater A: that's a bit unfair isn't it? I mean what do they have to do then ... if they're writing (Technology) to be able to do an 8 .. I mean if you're going to ... you're almost disqualifying them from being able to score highly

Rater C: perhaps then ... that's the wrong sort of question

Rater A: I think it's very clear... the message is very clear ... um ... it's well argued, there are some nice .. er... anaphoric are they? references ..."hence, aluminum" ... "which has a low density" ... and I ended up giving it an 8

Rater D: I do find it not so easy to judge things like ... the theme, logical presentation ... communicative effectiveness (pause) particularly as one ... in the other ones tends to think in terms of the relevance and accuracy of the arguments whereas we can't tell ... I suspect, for example, in line (4) "tough" is not the right word ... I would've thought "brittle" is the right word there ...

Rater B: I think there is a technical usage ... I mean I find it relatively easy to understand ... because ... I did do science to 'A' level and that makes it easier

Rater C: it sounds like textbook to me

Rater B: but it does ... I mean there are technical usages and I think tough is ... maybe ... I'm not quite sure
Rater D: that makes it difficult for us to assess it ... on the basis of what I can ... within those limitations ... I gave it a 7 ... I agree that there are a number of ... particularly ... er ... I mean, can you say "low hardness" and "high hardness" ... is that technical?
Rater A: No
Rater D: well how d'you know?
Rater A: because I've taught ESP and ... this area and I'm sure that's not right
Rater C: that last sentence is ...
Rater B: I think it should be "low hardness value" and ... there's little elements that ...
Rater D: that is a problem though ... if we ... he's obviously using "high density" which we do know
Rater B: but if you remember... I can't remember exactly ... they're in the Table ... now if you said at the top "hardness" and then with a number value in it ... you could expect a low hardness value ...
Rater D: oh I see ... I couldn't go up to an 8 because of the grammatical inaccuracies but ...
Rater A: I think probably I'd drop it to a 7 because of those, which I ... hadn't noticed
Rater B: that sentence there, look... "steel is strong and is therefore most suitable for this purpose" ... "although steel has ..." I mean there it would be ..."although it ..." although it has a high density it alsohas a high tensile strength" ... I mean it should be ...
Rater D: but then the fact that I understood most of it, I was so relieved that I .. um gave him a 7
Rater C: what sentences there are are not of a very high ... level of difficulty ... and they are ... jargon ... how else dwould you express these? ...I don't think ... I mean ... you can't say "hence" in the way that he's used hence ... um ... in fact ther's an error in nearly every sentence ... which er ... for some reason doesn't affect the ... er ... message ... the communication ... because I think it's almost note form
Rater B: but it's kind of a stylish error rather than a ... er grammatical error ... to a certain extent isn't it?
Rater C: I don't know
Rater B: I mean actually that's quite reasonable ..."for instance, the car body itself must be light in weight to improve the efficiency of the car" I mean there's nothing there lifted ... and it's ... er quite a nice formulation
Rater A: and then 'hence' that ... actually ah ... it seeems to me ... shows quite (inaudible)
Rater C: you can't after 'for instance' have 'hence'
Rater A oh yeh I think so ... I mean 'hence' goes quite nicely for me
Rater D: it's a bit odd to have it after a semi-colon ... it's a nice generalization opening sentence too

Rater B: yes, I mean ...
Rater A: I mean ...I think he knows how to write ... I mean not only has he got the informatiſpon there but he ...
Rater B: still ... you know comparing it with some of the other things we've seen .... I still think it's clearly 7ish

Writer Number 10

Rater B: It's a 6... mainly because it's very clear, I think, to read... the message comes across... there are occasional problems caused by the numerous repetitions of 'women' and 'woman' in not quite correct syntactic environments... um... but it's clearly better than a 5 because I don't think you have to strain very much to read it... there is an occasional strain... it's not a 7... because it's not really organised, it's not... well... very well structured...
Rater A: you've got to be joking... I started off... it had a certain spurious attraction... I started off with a 6-5-4 and then actually looking through it, it seemed that there was basically... you know, one reason repeated about 3 times in different ways for why... she'd like to read this book... um... the whole business of a woman's came up... the first time, and then... it came out again "moreover I think it is interesting to have an account of women by a woman" and then it came... er... in the sentence after that "Mary Size was the governor of women and hai the same women's sensibility" - it's just the same thing being repeated time and time again... and... for me.... on logic and argumentation dropped it... so I lost 6 and and had 5-4 and... grammar's not too bad... I gave it 5
Rater D: I started with a 4 and went on to a 4 and finished with a 4... when I got to "first", I was looking for "second"...
Rater A: yeah - me too...
Rater D: "for two main reasons, first"... der... der... der I wondered whether "moreover" was supposed to be "second" ...
Rater A: no it wasn't...
Rater D: ... but since it's the same point... or virtually the same point...
Rater A: yeah, and the "main interesting point" is again the first point...
Rater D: er... even the vocabulary was not brilliant... it's... you know... very limited... spelling's good, and it's neat...
Rater A: but - 'Marginal Writer'?
Rater D: I think it could be condensed into 2 sentences using all these words...
Rater C: I gave it a 5... I didn't give it a 4, I thought about giving it a 4 - or a 6 -, I had 4-5-6. 6: it definitely wasn't "Competent"... I thought it was a bit better them "Marginal". I know there's repetition but I mean it is...quite well disguised and it's um... she's expressed differently 3 times... this is every 'O' level candidate's dream...

Rater D: ... there might only be one argument... it depends how it's presented... but this is presenting one argument as if it was 3...

Rater A: it's suggesting that there are more and then... promising what it doesn't deliver

Rater A: ... and also... use of syntax... use of this "moreover"... I mean it's totally wrong... it's not a "moreover at all... that's... that's inaccurate..

Rater D: this smacks of pre-sessional... overkill...

Rater A: yeh... and then " the main interesting point" is the same point again ...you know... please make sure you've got all these things...

Rater C: perhaps she just misunderstands "moreover" and just means 'more of the same'...

Rater A: it does... she's been taught about it but is unable to use it - happens all the time in pre-sessional courses

Writer Number 23

Rater B: ...while there's a lot of good vocabulry... and lots of good strings, it is generally... it fits into that "the reader will have to strain on occasion to comprehend meaning"... because I did have to... because I found it... just going on and on and not coming to sort of a nicely rounded proposition or whatever... so I found it difficult to understand and I think it's on that communicative element that I have to... um...to go down to... say... 6... even though... there's some sort of good vocabulary and good strings...

Rater C: ...I went down to a 4 for the same reasons... because I did find it not just an occasional strian but a total strain... although the language is very impressive... well not the language structure but the.. the vocabulary is very impressive at first and er... you think he's saying something and... I don't think he is! - or she is

Rater D: the structure is pompous but it's clear... gives you the advantages first and then the disadvantage... the vocabulary's er... pompous if you like... bit... over-expressive - but I don't think you can penalise that - I mean, cosw e don't know what... the person... you... he's not given a target audience for this... I mean he isn't... you know... I can see that... if you specified a different target audience for the... you know... who you're writing for... he's not told who you're writing it for so I suppose it's... bit difficult... bit unfair... to penalise him on the type of vocabulary he uses... it is a bit... er... bombastic... but basically the message is clear... if tendentious... and the er... a lot of vocabulary... and a certain number of compensating... constructions...

Rater A: it seemed to me that I was getting the message... er... but it seemed to me that the argumentation, the organisation... was a

bit obscure at times and it was a bit difficult actually getting it through... and then because... I mean he... clearly has a nice grasp of the language, no linguistic inaccuracies that I could see...

Rater D: I can see the problems with it but the message is clear... I mean, he hasn't absolutely done it... he hasn't done it, particularly well in the last paragraph... I think most markers would be... seduced, like I have been... by the bombast

Rater C: well, that just put me off entirely - I thought, anybody who can write that sort of thing...

Rater D: but that's just a cultural thing (inaudible)

Rater C: I don't see why being a porter could help to explain many diseases and procedures, which is what he's suggesting... "moreover they are a continuation of their previous skills" - I'm sorry - "moreover, they now understand and are able to explain many diseases and procedures"- what on earth has that got to do with

Rater D: (interrupts) I think in the context in which they 're working... where doctors are often very much detached from what actually goes on there...

Rater C: oh I see...yes - yes

Rater A: (interrupts) "the auxiliary nurse"

Rater D: perhaps he's emphasizing the "able to explain", that is, that you've got the touch of the common people... but it's obscure, isn't it?

Rater B: because of the communicative element there, I think it's difficult to give it a 7...

APPENDIX H

UNIVERSITY OF EDINBURGH
Institute for Applied Language Studies

21 HILL PLACE • EDINBURGH EH8 9DP • Tel 031-667 1011 Ext 4592 & 4596 • Telex 727442 UNIVED-G

Liz Hamp-Lyons

OVERSEAS DIPLOMA/MSc. STUDENTS AT UNIVERSITY OF EDINBURGH

SUPERVISOR QUESTIONNAIRE

Your Faculty_____Your Dept._____

A. How often do students on Dip/MSc courses have to perform the following writing tasks?

1. write short introductions or connecting sentences in numerical calculations or mathematical arguments during
   a) coursework: often sometimes never
   b) examinations: often sometimes never

2. write short connected answers to questions demanding a narrow response (ie. where the question states the limits and nature of the response required) during
   a) coursework: often sometimes never
   b) examinations: often sometimes never

3. produce extended writing (ie. continuous writing longer than single paragraphs) during
   a) coursework: often sometimes never
   b) examinations: often sometimes never

4. produce any other types of written work (please describe)

B. How often are Dip/MSc students asked to do the following in their written work?

1. make a list of concepts, ideas, or events _____
2. summarise readings or lectures _____
3. compare or contrast one concept, theory or idea with another _____
4. apply models, principles, or generalizations to a new situation _____
5. argue a position _____
6. analyse (break down information into constituent parts) _____
7. synthesize (produce something of their own from what has been studied or observed) _____
8. evaluate using internal evidence or external criteria _____
9. other (please describe)

C. How often do you find students on Dip/MSc courses who have these defects in their writing?

1. grammatical errors _____
2. restricted grammatical choices _____
3. inappropriate grammatical choices _____
4. vocabulary errors _____
5. restricted range of vocabulary choices _____
6. inadequate understanding of the subject _____
7. inability to express themselves clearly _____
8. poor arrangement and development of written work _____
9. poor spelling _____
10. poor punctuation _____
11. poor handwriting _____
12. untidiness _____
13. other (please describe)

D. How much importance do you attach to these features of writing in the written work of Dip/MSc students?

1. grammatical accuracy _____
2. variety of grammatical structures _____
3. appropriateness of grammatical structures _____
4. accuracy of vocabulary _____
5. range of vocabulary _____
6. subject content _____
7. clarity of expression _____
8. arrangement and development of written work _____
9. spelling _____
10. punctuation _____
11. handwriting _____
12. tidiness _____
13. other (please describe)

E. What proportion of your examination requirement for Dip/MSc students is in the form of questions to which the students must write answers?

F. What design factors do you keep in mind when preparing examination questions, apart from the actual content to be examined?

G. What criteria do you use for marking examination answers, in order from most to least important?

THANK YOU FOR YOUR HELP.

STUDENTS' ANSWERS:

APPENDIX G: CHAPTER 5 SECTION 2

APPENDIX I: CHAPTER 6

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX 9

WRITER #3 Band [ ]

M2 (WRITING) TEST TECHNOLOGY MODULE

...

...

st

Question 1: Look at the table on page 5 of your Source Booklet which describes certain characteristics of steel, iron and aluminium. Discuss the advantages and disadvantages of particular metals for a particular purpose with which you are familiar (e.g. building construction or vehicle manufacturing). (Write 15 to 20 lines).

In vehicle manufacturing, the choice of materials for making certain parts of a car is extremely important; the properties of a particular metal must meet the mechanical and structural specifications which could serve the purpose. For instance, the car-body itself must be light in weight to improve the efficiency of the car; hence aluminium, which has a low density of 2650 kgm^{-3} is most suitable for this purpose. There are of course other materials like glass-fibre, plastics or even carbon fibre which are extremely light in weight in comparison with aluminium but the overall performance of aluminium outweight the others in that it has a reasonably high hardness, high modulus of elasticity E and is a ductile metal which brings about its success.
Another example like the wheel-axle of the car must be very strong but not tough. Steel is strong and is therefore most suitable for this purpose. Although steel has a high density, it has a high tensile strength. Aluminium has a far too low hardness and iron has a low E value, these are the reason why they have been ruled out.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

473

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX 9
WRITER #4

GENERAL ACADEMIC MODULE

Test Ce ..... Date .....

Candida .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 5.I. of your Source Booklet. If yc had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have knou* state which one you would choose ~~to~~ write 15 to 20 lines giving the reas for your choice.

If I had to read either Henry's 'Who lie in Gaol' c Size's 'Prisons I have known', I would prefer to choos the second one. 'prisons I have known'. Because:

1. It is a personal account. personal account more reality, more attractive.

2. It has been quet a long time for she wrote the Forty-seven years for a worker. there must be ra lot experience. so the book must be in good detail concern the prisons.

3 She was the first governor of open prison. She must be very intellenge. and did very work in her posi

4. she had been in many prisons as a governor. so she must have many knowledge about the prisons an how to rule the prison.

In addition to that she have an excellent accoun st the open prison. so I would like to choose thi book

474

IF NEGESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

APPENDIX G
WRITER #5

Question 1: **New scientific processes often meet with opposition because of the pollution they cause to the environment, an example of which is referred to in Section 2 of your Source Booklet. As a scientist, how would you defend the continual use of such potentially harmful processes? (Write 15 to 20 lines).**

Nowadays many people complain because of new scientific processes. They assert that these processes cause pollution and consequencly the environment is destroyed. Indeed, espesially the last years, the problem of pollution is very big. On the other hand, we must admit that scientific processes cause the progress of our life consequences. In addition, the life expectance has been increaced, and land efficiency has been developed better by means of these processed
For example, agricaltural machines cause air pollution because they operate with petrol On the contrary without these machines, many people could be died because of malnutrition, even starvation
As a result, we can say that the usage of factories, or plants, or other machines which cause pollution is strongly associated with our life, although the danger of pollution continue to exist.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

APPENDIX 9

ENGLISH LANGUAGE TESTING SERVICE

WRITER # 6

Band

M2 (WRITING) TEST

MEDICINE MODULE

Test Centre ... Date .

Candidate's .....

Answer BO answer, a ease allow 25 minutes for the first

Question 1: Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines).

In my opinion the benefits which a Doctor could derive from a part spell as an auxiliary nurse, porter or technician greatly depends a the character of the Doctor.

It might help some to ~~get a better~~ understand the patients and their problems better and it might also improve their future relationships with nurses and other senior members of the hospital administration.

It could make the Doctor more considerate of the problems nurses and technicians face.

On the other hand some Doctors might come out of this experience completely unaffected.

It could even have an adverse effect on some by not only putting them off some specialities like psychiatry for example but ~~completely~~ may even make them lose interest in Medicine as a whole!

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX 9

WRITER #7 Band [ ]

M2 (WRITING) TEST MEDICINE MODULE

Test Centre ....

Candidate's Name .. ..

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines).

The medical field is very wide, and working in one of its branches, is always a ~~benefit~~ benefit, even if you want to move to another branch. Working as a paramedical personel is definitely a very good experience for Doctors to get before being a Doctor, that is because being in the service in whatever status might be, gives you confidence in yourself, gets you acquinted with everyday work, and ~~you develope the~~ you get yourself into the habit of being involved in the profession, you feel home the first day you join your hospital as a student, The medical terms would be much easier for you to pick up. learning to deal with the patient would be faster. In all almost every thing is familiar to you which makes your more at ease in every way. But working as a porter or otherwise other than being a stude or Doctor is a waste of time in away, because the time you spend in doing such things could be more benefiting if it's spent in studying or learning the real medical profession.

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX G

WRITER # 8

Band [ ]

M2 (WRITING) TEST MEDICINE MODULE

Test Centre ..... Date .....

Candidate's Name .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines).

I worked in hospitals for several years as a resident doctor, I did a year a resident internship. I had a clinical experience as a student. During that time I have the follow-up notes.

It is essential to have a good medical knowledge because we can deal with the patents without adquat medical knowledge.

To work as a nurse or porter facilitat for as to have some experience how to deal with patient to have some idea who the treatment be complilly. because we give the patient drug. we must know How this treatment reach the patient

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

478

TURN OVER FOR QUESTION 2

APPENDIX G
WRITER # 10

Answer both questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 5.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have known*, state which one you would choose and write 15 to 20 lines giving the reasons for your choice.

I would certainly choose Size's "Prisons I have known" and that for two main reasons: first, because I think the experience of Mary Size is almost unique: she is at the same time a woman, an author and the first Governor of Askham Grange "open" prison. There are very few accounts like that on life and administration in the likewise prisons. Moreover, I think it's interesting to have an account on women and girls life and problems in prison written by a governor who is also a woman. The main interesting point is to see the reactions, the discussions, the problems which can be caused by the fact that Mary Size was the governor of women prisons and had nevertheless the same women sensibility than the women who lived behind the walls of the prisons she managed..

478(a)

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

APPENDIX G

WRITER # 11

Question 1: Refer to the bibliography in Section 4.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have Known*, state which one you would choose and write 15-20 lines giving reasons for your choice.

I would rather Size's "Prisons I have Know".
The reasons for my choice are that in this book I will find broader information about different Prisons, including Askham Grange, the open prison for women. Also, because of it had been written by a educated person, I will find it less parcial to emotion. Of course, I recognise that in a way, this book could be no realistic as the Henry's one. Here for instance I might find better information about treatment and inmates, but from a parcial point of view, the prisoner. In conclusion, the best thing would be to read both books so that to obtain full information about the subject. One of them will illustrate you from a goverment point of view and the other one from the prisoners.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

APPENDIX G
WRITER # 13

on 1: **New scientific processes often meet with opposition because of the pollution they cause to the environment, an example of which is referred to in Section 2 of your Source Booklet. As a scientist, how would you defend the continual use of such potentially harmful processes? (Write 15 to 20 lines).**

New scientific processes often result in pollution such as air pollution, water pollution, sound pollution, etc. Thus, as a scientist, he must always have the idea that how to minimize the pollutions that the new scientific prusses bring about. New methods, and amendments must be used in order to reduce or prevent the pollutions. Air pollution can be prevented or be reduced its harmful by first liquefied the harmful gas, and then pour into some safety cans, these cans are then completely sealed and use for reclamation. Many waste products in industrial areas which are harmful are poured into the sea. Chemicals must be added to these waste products before pouring to decompose or form compounds with the harmful substances.

479(a)

ESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

WER FOR QUESTION 2

APPENDIX G
WRITER # 14

answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 5.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have known*, state which one you would choose & write 15 to 20 lines giving the reasons for your choice.

I would choose Size's Prisons I have known, because I have seen so many films about prisoners and I have read accounts from prisoners about the prison and other prisoners. Now I am interesting to read

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX G

WRITER # 15 Band [ ]

M2 (WRITING) TEST GENERAL ACADEMIC MODULE

Test Centre ... Date ... .

Candidate's I .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 5.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have known*, state which one you would choose and write 15 to 20 lines giving the reasons for your choice.

I choose Henry's 'Who lie in Gaol'

In the life of prison, there are 2 sides. One from the prisoner's side, the other from the governor's side. Both sides are important and we must look at the life from both sides. However, since the pison is basically for the prisoner to make themselves ready to go back to real world, not for keeping the criminals off the society, nor for keeping the prison in order, we should consider the side of the prisoner first.

There is a well-known psychological experiment which try to simulate the imprisonment in the laboratory. Perfectly normal people were randomly assigned either as prisoners or guards. They played the role in the laboratory for about a week. The result was shcking. Despite the clear understanding of the situation, that is it was an experiment and they were only playing the roles, the subject behaved very much like prisoner and guard. Guard became more and more agressive and demanding - sometimes very cruel. Prisoner became timid, obedient and felt helpless It was clear that the situation of imprisonment itself has in evel characteristics and develop inhuman relationship among the people there.

The book written by ex-governor of the prison could be very well written and good description of what the objective life is like in prison. But we must know that in spite of her best wishes to help the prisoner, her eyes were the eyes of the governor.

Thus, if I have to read one of these books first, I would read a book written by the person who was actually in prison first.

:R SIDE OF THIS SHEET 481

The other book should be read next.

APPENDIX G

WRITER # 16

Question 1: Refer to bibliography in Section 4.1. of your source booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have known*, state which one you would choose and write 15 to 20 lines giving the reasons for your choice.

I'd like to choose the Size's 'Prisons I have known'. Because I'm not quite sure about prison and cannot imagine the inside of it at all. From the view point of 'architecture', my profession is an architect, I would like to know about the conditions of prisons; that is what kind of persons are there, what kind of rooms are there, what kind of facilities are there, what is their daily life and so on.

After that I'd like consider the good physical conditions for prison.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

APPENDIX G
WRITER # 17

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 5.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have known*, state which one you would choose & write 15 to 20 lines giving the reasons for your choice.

I would choose Size's "Prisons I have known" for the following reasons.

~~Firstly~~ though it would be my first attempt to study the causes and treatment of crime and prison conditions today and in the past, I have already accumulated a certain amount of knowledge about the issue ~~from~~ from ex-prisoner's point of view. ~~Therefore~~ ~~therefore~~. As the author of this bibliography suggested that ex-prisoners work could not illustrate the whole picture of the issue, I am supposed to read Size's works.

~~Taking into account that:~~

1. My previous knowledge ~~on this issue~~ is biased in favour of ex-prisoners.
2. ~~I should be fair in analyzing the issue from the~~ both prisoner and

Because it ~~was~~ was written by many as governor of prisons and Borstals who were serving for prisoners, I expect it to give me a different point of view from that I used to.

482(a)

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

APPENDIX 9
WRITER #19

ENGLISH LANGUAGE TESTING SERVICE

Band

SOCIAL STUDIES MODULE

Test Cent .....

Candidate .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 4.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have Known*, state which one you would choose and write 15-20 lines giving reasons for your choice.

FIRST I WISH TO STATE, THAT THE OUTCOME OF MY CHOICE DEPENDS ON WHAT KIND OF MATERIAL I AM LOOKING FOR. OBVIOUSLY HENRY'S BOOK CONTAINS THE CONVICT'S PERSPECTIVE ON THE PRISON SYSTEM, WHILE SIZE SUPPLIES YOU WITH THE STAFF'S POINTS OF VIEWS. MY CHOICE BETWEEN THEM DEPENDS ON THE PROBLEM I WISH TO STUDY. IF MY MAIN INTREST IS IN PRISON ADMINISTRATION OR THE SOCIAL AND IDEOLOGICAL FEATURES OF THE PRISON STAFF, I WOULD NATURALLY CHOOSE SIZE'S BOOK. IF, ON THE OTHER HAND, I AM MORE INTERESTED IN THE ACTUAL EXPERIENCES OF CONVICTS FROM WITHIN THE PRISON SYSTEM, THEN HENRY'S DESCRIPTION WOULD BE MY CHOICE.

THIS IS, HOWEVER, MERELY A DISCUSSION OF PREFERENCE, OF RELATIVE WEIGHT. IN AN ANALYSIS OF A SOCIAL SYSTEM LIKE A PRISON YOU CAN NOT WORK EXCLUSIVELY WITH MATERIAL FROM ONLY ONE SIDE OF THE BARS. THE VIEWS OF STAFF AND CONVICTS MUST BE CONFRONTED AND ANALYZED IN ORDER TO OBTAIN A MORE COMPLETE, HOLISTIC PICTURE.

482(b)

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX G
WRITER #21 Band [ ]

M2 (WRITING) TEST LIFE SCIENCES MODULE

Test Centre .

Candidate's N

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Section 4 of your Source booklet deals with the Green Revolution. Drawing on your own experience, discuss some of the advantages and disadvantages of the introduction of modern farming techniques. (Write 15 to 20 lines).

Rough Draft:
Modern farming techniques Introd. (3)
A ADVANTAGES
1 Possibility of feeding a larger population on a smaller area < quantity / time required to grow / less waste
2 Greater variety of crops could be grown. (6)
B DISADVANTAGES
1 Uncontrolled use of chemicals could endanger human population
2 Unemployment as machines do most of the work (6)

MODERN FARMING TECHNIQUES

With increasing global population in a limited and obviously not expanding universe, man has had to adapt to the situation. One way he has done this is to invent modern farming methods.

The merits of the new techniques are many, but the most important of them would be the possibility of feeding a larger population in a small area as quantity harvested would be higher, and it would take less time to do it. Further, a greater variety could be grown as artificial conditions could be introduced.

However, as the new techniques involve use of many chemicals, the one who should have benefited, man, could be harmed by uncontrolled use of the pesticides and fungicides among others. It is worth noting, also, that there would be less employment as machines would do most of the work.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

482(c)

APPENDIX G
WRITER # 23

:ion 1: Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines).

The advantages and disadvantages that a medical student or a doctor could have had by having some previous experience of hospital life depends upon their respective duties and the skills they had to employ. Thus, those who come more frequently in contact with patients have a better chance of establishing rapport with them provided the doctors do not feel themselves standing on a pedestal. Secondly, doctors with previous paramedical experience are well aware of the hospital organisation and those previously practising technical skills adapt nicely to situations which are a continuation of their previous skills. Moreover, they now understand and are able to explain many diseases and procedures, which are nothing new to them.

I think that the only disadvantage from a previous hospital experience as paramedical personnel is not getting used to the idea that one has to work with a broader perspective, maintaining the humility which one has cultivated previously by occupying a less rewarding position of a paramedic.

:CESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

:OVER FOR QUESTION 2

482(d)

APPENDIX I

GENERAL ACADEMIC

You have 30 minutes to answer this question. Answer on the front of this sheet.
You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words.
You are not expected to show specialist knowledge, but your answer should be relevant.
Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Explain why the 'green revolution' of high technology in food production has created serious social problems in India. Refer to pp 7/8 in your Source Booklet.

The new seeds which had breaden by Dr. Norman Borlang have to have regular suppiles of water – so only irrigated fields can be planted. But India has not irrigated fields so much. And the new technological method needs a large scale and generally with the help of machines.

In a words, the green revolution required a sizeable investment, so the goverment supplied cheap credit for tractors, expensive seed and fertilisers to be bought it was the well-off farmer who had more than enough land to provide for his family's food needs and could risk giving the new seeds a whirl. It was the well-off farmer who had the irrigated land. And it was the well-off farmer who had large enough fields to make a tractor a worthwhile asset.

But small farmers with their regular crops found that market prices had been driven down. Because the harvests of the well-off farmer had been dumped.

In a word, the green revolution made the gulf between the village rich and poor wor wide.

4-83

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX I

Band

~~WRITING~~ TEST LIFE SCIENCES MODULE

Question 1: Section 4 of your Source booklet deals with the Green Revolution. Drawing on your own experience, discuss some of the advantages and disadvantages of the introduction of modern farming techniques. (Write 15 to 20 lines).

We are impressed, in developing countries, of the modern farming techniques because we read many reports, experiments and different works which were done on this field and showed good result.

When we practised introducing a new variety of seed or fertilizer or pesticide we faced many difficulties although they have many benefits.

When introduced seeds are grown in our conditions they reacted differently from their first locality. For this reason they need a number of years to be adapted to the new conditions. We applied fertilizers to plants they had good results at the beginning but that yields became lesser and could not be used economically for producing more yield.

New pesticides were used to protect plants. They gave good effect, but after certain number of years pests had done many damages on plants and applying pesticides was useless.

We can conclude that applying modern farming techniques has some disadvantages for a certain limit but we still use it because of the advantages which we get.

484

EXP 18

LIFE SCIENCES

You have 30 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Explain how you would eliminate an undesirable genetic characteristic from a herd of cows. Refer to pp. 3/4 in the Source Booklet.

When we try to eliminate any undesirable genetic characteristic from a herd of cows, we first have to know the nature of that particular genetic characteristic. If the one in concern be associated with a single genetic locus, we could simply apply Mendelian laws. What we have to do is just to determine (or estimate) the genotype of the offsprings and to cull. If the genetic characteristic which we want to eliminate envolves many genetic loci, as it would be the case normally, we could try to estimate how complex the genetic system is by carefully analysing the pedigree. What we would actually do might not be different from the first case, just to cull the offsprings which express the worst of the character which we want to eliminate. The problem here is that the life time of cattle is as very long as 20-30 years and therefore it takes long time to do the tasks. In both cases, it is important to know the statistics of the population in concern not only of the characters which we want to manipulate but also of the characters which we could link genetically to the former. Since artificial insemination is the household technique now-a-days, very careful selection of the genetic sorce from various stocks of bulls might be expected to do it job sometime very well. All things however depend on the nature of the genetic characteristic in question.

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX I

Band

M2 (WRITING) TEST PHYSICAL SCIENCES MODULE

Question 1: **New scientific processes often meet with opposition because of the pollution they cause to the environment, an example of which is referred to in Section 2 of your Source Booklet. As a scientist, how would you defend the continual use of such potentially harmful processes? (Write 15 to 20 lines).**

Even though I am a scientist, I strongly consider the opposition to new scientific processes as a healthy action. I think that man's new experiments should compromise their experiences and the environment, in order to avoid damages to the present and future nature. Basically, no man or nation has the right to, in name of a scientific progress, destroy their own habitat - the only one we have now.

Fortunately though, I believe that man can find better ways in order to garantee the environment and mankind preserved. Indeed, I also believe that new scientific processes can be done just to improve the quality of the life on Earth, in despite of all economical interests involving the man's action all around the world. Even though someone can find this thesis completely utopic, I really trust it.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

486

TURN OVER FOR QUESTION 2

EXP 31 , APPENDIX I

PHYSICAL SCIENCES

You have 30 minutes to answer this question. Answer on the front of this sheet.
You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words.
You are not expected to show specialist knowledge, but your answer should be relevant.
Although grammar, spelling etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Discuss some of the ways in which air pollution can be reduced.
Refer to p.3 in your Source Booklet.

At first the Source Booklet should induce us to think about a Large-Term plant for separation of ~~Air~~ air into its constituents. As suggested there, we could liquefy a large amount of air by cooling as different constituents have different vaporisation temperature. In this way, the undesirable part such as carbon dioxide or sulphur compounds should be eliminated. At least by now, I don't think it should be the most intelligent (or even cheapest way) to reduce the air pollution. We already have available a natural way - the forests. Just for keeping the ~~forests~~ already existent forests and/or making it grow near polluted areas, we could contribute for a more enjoyable life in our planet. Besides, severe policies against factories ~~outputs~~ outputs are strongly necessa

Anyway, the only enemy one's should foresee is the own man, all involved in getting as money as possible that don't have time to ~~first~~ think about such a simple and primordial question.

EXP 35

APPENDIX I

MEDICINE

You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. If you were a general practitioner in West Africa looking for information on how to reduce the mortality rate from bilharzia in your area, which of these books would you refer to and why? Refer to p.9 in your Source Booklet.

If I were a general practioner in West Africa looking for information on how to reduce the mortality rate from bilharzia in my area I would refer to these books :-

The first books I'd refer to is the international Journal of epidimiology 1976, 5 NO.5, This is because there are two articles, in this book, from which I can know the prevalance and the morbidity of the disease the two articles are, A household morbidity survey in rural Africa and surveillance of communicable diseases in tropical Africa. It is an important thing to know the magnitude of the problem before do any solution.

The second book I'd refer to is the Tropical Doctors 1976 6 to read the article, the challenge faced by the medical profession in tropical developing countries. Because, in this article I'll find the ways how the doctors face the diseases in the developing countries, the diseases which are mainly communicable, and how they manage and eradicate the diseases.

Consequently, by adopting the two books to read three articles I think I'll find the possible way of reducing the mortality rate from bilharzia.

488

EXP 36
APPENDIX I

GENERAL ACADEMIC

You have 25 minutes to answer this question. Answer on the front of this sheet.
You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words.
You are not expected to show specialist knowledge, but your answer should be relevant.
Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Explain why the 'green revolution' of high technology in food production has created serious social problems in India. Refer to pp 7/8 in your Source Booklet.

Actually, the 'green revolution' has brought a great advantage. However, we should pay attention to another aspects of it. The new varieties of seeds need to be pampered or they sicken and die, and have to have regular supplies of water. As a result, just irrigated fields can be planted. This means they need expensive artificial fertilisers.

Moreover, the green revolution caused a change between the rich and poor farmers. In other words, it resulted in a discrimination. Because only well-off farmers can buy the new seed and only they can take a risk in doing so. Anyway, the poor farmers still remain poor.

In my opinion, what is called 'green revolution' in India will be called 'CATCH 22', which will be always in backfire.

APPENDIX I
EXP 42

SOCIAL STUDIES

You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Describe the effects of the fall in death rates in Western Europe. Refer to pp 3-5 in your Source Booklet.

It is seen that in Britain and western Europe, the exp
tation of life at birth is about 70 years. This is
mainly due to the development in medical kn
wledge and sanitation system. A second factor
involved is the benefit of advanced agriculture
and industry which gave a better standard of
living. About two centuries ~~are involved~~ were taken in
reduction of death rates in western Europe.
At the end, it is concluded that application of
scientific and medical advances has resulted in
the death control in western Europe.

I think that it is very difficult to answer this question. The reason is that, as ethics is a subjective concept in each particular community, the meaning of the "sins" is varying a lot.

Furthermore ethics refers not only to the community but also to each one of us seperately. So, the action of committing a crime has a different meaning to each one of us and to each one community. Subsequently we have to judge each case according to the custom that is according to the way of life of each particular community.

For example: it is generally believed that killing is the most terrible "sin". The people supporting it think that we can not do something like that because life is given by the God and we have not the right to kill someone irrespective of the reason. But these people can stand the everyday starving of thousand of people of the Third World while they know that is essentially a way of killing all these people by the developed countries.

It is also believed that stealing belongs to the biggest "sins". But who can support that some people (very few) must have too much and all the rest must starve? Someone of course, here, can refer to the extremist groups who (it is believed) steal and kill without reason. I am not for these groups, but we have to take in account that any unfair situation produces resistance. And I really believe that the society is bad structured. So, if we see the things from the other point of view may be we can justify them.

Regarding other "sins" like not honouring parents or bad manners, I think, that we have to refer again to the structure of the society (education et.c). If we can not afford some elementary things to all the members of the community it is unfair to ask a lot of them.

Anyway, I believe that everything which gives happiness to someone must be done without of course damaging or damage someone else, because freedom of each one must stop when the freedom of his neighbour begins.—

EXP 53
APPENDIX I

LIFE SCIENCES

You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Explain how you would eliminate an undesirable genetic characteristic from a herd of cows. Refer to pp. 3/4 in the Source Booklet.

In a herd there may be different types of undesireable genetic characteristics. For example, the ~~breeds~~ progeny may differ in colour or milk yield may reduced than earlier generation and so on.

To avoid these ~~genetic~~ undesirable genetic characteristics, the breeder should have experience on phenotypic character of the progeny. With this experience he can made proper selection ~~before~~ among the breeding cattle ~~by observing its~~ before breeding.

He should observe the yield per lactation and the annual yield of the parent and grand parents of the breeding animals.

Following these procedure he can a breeder can ~~avoid~~ eliminate undesireable genetic characteristics from a herd.

EYP 60

APPENDIX I

1. Peoples A man's belief in sin ~~has~~ is one of the most serious sins. Because when he does ~~anoth~~ something, he immediately thinks whether or not there is any good effect out of his works. If he feels that there are some evil effect out of his works, it is his sin which leads some moral depression upon him.

2. Another serious sin is ~~the~~ the regretness of people. A man having done something should not regret that ~~they~~ he is wrong. For ~~ever~~ If he always thinks that what he has done is wrong, it keeps him unhappy in most of the time. That's why the rich people are found to be unhappy in comparison to the poor.

3. Another serious sin is to believe in God. and. Because nobody can gives any proper definition about about God. But some people are regular-worshipers. They have faith in God and consequently thinking about God they sometimes show their strict attitude towards others. Sometimes such strict attitude brings potential harmfulness to them.

4. Another serious sin is the question of honesty People think that honesty is ~~the~~ that should be the most important quality of human beings. But in practical life it is found that most of the people are dishonest. Because dishonesty has put a man in trouble dishonesty is also another serious sin.

APPENDIX I

GENERAL ACADEMIC

EVP 74

You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

1. Explain why the 'green revolution' of high technology in food production has created serious social problems in India. Refer to pp 7/8 in your Source Booklet.

Unesco Office of Statistics attempted to update to 1980 the estimate of the world literacy situation in 1967-1969

At that Literacy tell us too many problems that the 1970 estimated world figure of 783 m. million illiterate of decrease in illiteracy been maintained.

However the ever-rising of increase of population still cause the rate of increase in the ~~over~~ number of illiterates to rise.

The green revolution movement is begin in India in 1970.

The green revolution in India reached its high-water mark in 1970/71 when a record-beating crop of over 100 million tons of foodgrain were harvested.

But the arhievements have turned sour.

I think green revolution is very important revolution for farmers and poor people

We must connect many the other theris and social Society.

EYP8+

APPENDIX I

ENGLISH LANGUAGE TESTING SERVICE

Band [ ]

SOCIAL STUDIES MODULE

Test Centre .... Date .. ..

Candidate's ....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: 87 Refer to the bibliography in Section 4.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have Known*, state which one you would choose and write 15-20 lines giving reasons for your choice.

At this moment, I am not really intersted in prison's problems. Surely, Size's Prisons I have known is a well documented book and I might learn a lot about prisons reading it. I would like to know about problems of "open prison". However, if I have to choose only one book to read, I prefer to read Henry's Who lie in Gaol. I expect it to be like a novel. I would like to find an easy book to read Aditionaly, I think I will find interesting to hear of personal experiences of a prisoner. Generaly, one has opportunities to know opinions from the police, authorities, etc, but rarely there is opportunity to know what prisoners think. I feel I have a professional interest in this book. As a psychologist I am interested in knowing emotions, reactions, feelings of a man in such a hard situations.

494(a)

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

APPENDIX I

ENGLISH LANGUAGE TESTING SERVICE

Band [ ]

Test Centre . .....

Candidate's N .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Refer to the bibliography in Section 4.I. of your Source Booklet. If you had to read either Henry's *Who lie in Gaol* or Size's *Prisons I have Known*, state which one you would choose and write 15-20 lines giving reasons for your choice.

93

if I want to choose one of these books which are mentioned above, it depend on what I really I want to study either to build up a background or to collect a date to make a research. If I have to to choose one, I will choose the book which is written by Sizes, as he had an experiences and he was dealing with different type of prisons, so any one intrested to write something about the real life of prison, this book will provide him with a real information, But think it doesn't mean, the researcher just depend on one reference it's this kind. he should have collect his date from different resources to make a good decision.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

494(6)

APPENDIX I

LIFE SCIENCES

You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling, etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Explain how you would eliminate an undesirable genetic characteristic from a herd of cows. Refer to pp. 2/3 in the Source Booklet.

I would like to explain at this point, that characters that neither the bull or the cow can transmit to the calves are given by two types of genes: (1) DOMINANT and (2) RECESIVE.

In case this undesirable genetic characteristic is located in the latter group of genes it could not appear in next generation of calves because they may be "cover" by the former group, that is the DOMINANT.

So the DOMINANT genes are the most important group and they need to be treated because they are the stronger genes and always appear in the next generation.

- a result, if the undesiderable genetic characteristic is located in DOMINANT genes, the cows which contain this genes must be separated from the heard to eliminate the problem.

Otherwise, these undesiderable characteristic will be carried from one generation to the next generation of calves.

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX I

Band

M2 (WRITING) TEST

MEDICINE MODULE

Test Centre ....

Date .....

Candidate's Name .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: Section 5 of your Source Booklet states that a doctor can benefit from having had some experience of hospital life as a porter, technician or auxiliary nurse. Give your own opinion of the advantages and disadvantages of such experience. (Write 15 to 20 lines).

Medicine is a hard Job and very stressful - so if you
wont to be a doctor you must accept this fact and try
to best to be a good doctor.

In my experience on a doctor I will say that, the doctor
should know every thing about this Job which he was
trained. What I mean here, things which he might
think that there are less important for his examination
and prescription of the curative drugs. The doctor
It is very important to know all about the
work of the para-medical staff as nurses and the clinician,
and the non-medical staff as the porters and receptionists
because he will need them everyday. And especially
in a general practice where he may work alone as a
single-handed doctor or as many accessories like Emergencies.

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

In my opinion, there are very much disadvantages in this experience. I think the only disadvantage is the medic— student may change his mind towards a certain branch of medicine because of a bad situation he was put on experi. in the field of psychiatry.

It not a true idea that doctoring is depend only on the doctor but sometime the nurse or other member of the staff can deal with the problem even better than the the doctor. Experience is needed in many works and for the doctor it is important to know his job and the other jobs in his field.

Question 2: Refer to Section 3 of your Source Booklet. You have sat in on the interview between the doctor and the patient, Mrs. Jellicoe. You wish to discuss the case with the doctor before Mrs. Jellicoe's next visit. Make a few notes about her main symptoms and their probable causes and the action taken by the doctor.

The patient named Mrs. Jellicoe has had a problem related to her nerves which is mean she in her had a psychiatric condition. Her problem mainly because of a state of degression which developed after the operation in her breast and the home problems with her husband. She still worried about that operation and she is thinking of her house and this mates her sik and asking for treatment. It is a good idea to start with some sedatives tablet to makes her calm and after sometimes may be for two week. the doctor can see her again and re-examin her to find a suitable treatment for her.

EXP 120

APPENDIX I

PHYSICAL SCIENCES

You have 25 minutes to answer this question. Answer on the front of this sheet. You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words. You are not expected to show specialist knowledge, but your answer should be relevant. Although grammar, spelling etc. are important, we are most interested in your ability to orgânise and communicate information and ideas.

Q1. Discuss some of the ways in which air pollution can be reduced. Refer to p.3 in your Source Booklet.

Pure air is consits of Nitrogen 78%, oxygen 21% & argon 1% addition to other inert gases and CO_2 about 0.03% & to maintain these ratios constant we must control the side production of other gases which are dengerous for the life & for that we control this by

(1) In factories which produce agases as aside product We a could absorbe these gases by chemical Processes to convert these gases to another compounds which are usefull.

(2) By using the filters & not to allow these dengerous gases to go to the open air & then condence these gases & convert them to compounds

(3) In factories we try to use amethods that produce ~~the~~ agases which usefull to put them in containers such CO_2, O_2, N_2, H_2 etc.

④ The continuous uses of O_2 by life cycle & by manufactures this will reduce it's porcentage in the air & for that we must increase ~~the~~ ~~the~~ ~~Plants~~ the number of trees & to cultivate the land in a wide distances & for that we can get more O_2 & also to filterate the air from dust

⑤ Some compounds in air are may be condensate & be usefull

ENGLISH LANGUAGE TESTING SERVICE

APPENDIX I

Band

M2 (WRITING) TEST PHYSICAL SCIENCES MODULE

Test Centre ... .

Candidate's Name .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: New scientific processes often meet with opposition because of the pollution they cause to the environment, an example of which is referred to in Section 2 of your Source Booklet. As a scientist, how would you defend the continual use of such potentially harmful processes? (Write 15 to 20 lines).

120

The continual use of such potentially harmful processes will change the percentages constituents of environments & for this we must try to minimise the side productions such as dust, soot & sulphur compounds which come fra factories.

To controle these products we can try to absorbe them by many modern methods to get agreat useful by using them to other product as aprimary constituents & not let them to ha in our en· onment. And we can also find other methods by which we will not get much side harmful prod, also we must bild the factories away of centres of towns & to increase cultivations around towns & big cities to decrease · dust & other impurities to harm the environment

IF NECESSARY YOU MAY CONTINUE YOUR ANSWER ON THE OTHER SIDE OF THIS SHEET

TURN OVER FOR QUESTION 2

496(6)

CAT 124

APPENDIX I

PHYSICAL SCIENCES

You have 25 minutes to answer this question. Answer on the front of this sheet.
You may use the back of the sheet for rough notes, which will not be assessed.

If you use information from the Source Booklet, put it in your own words.
You are not expected to show specialist knowledge, but your answer should be relevant.
Although grammar, spelling etc. are important, we are most interested in your ability to organise and communicate information and ideas.

Q1. Discuss some of the ways in which air pollution can be reduced.
Refer to p.3 in your Source Booklet.

Pollution is one of the greatest problems facing man kind today. Industury is to blame for some of the pollution. Facturies and cars exhusts are the main reasons of air pollution.

To reduce the air pollution we can move the facturies from the citys or the facturies should be ordered to remove the poison from their exhusts before they exchang. (also remove the dusty, we can use anti pollution in the air
The trains and plans -- ships ... are also from the reasons of air pollution.
We need to grow and build more garden to avoid the air and make the air fresh.

APPENDIX I

Q2. Read the text below and then answer the question under it. You have 35 minutes altogether for reading and writing. Use the back of the Q1. paper for notes if you wish.

British still believe in sin, hell and the devil

by JUDITH JUDD

MOST Britons still believe in the concept of sin and nearly a third believe in hell and the devil, according to the biggest survey of public opinion ever carried out in the West.

Britons have a stricter moral code than their fellow Europeans, especially about sex under the age of consent, fiddling the dole and keeping money they have found. But they are more permissive about euthanasia and failing to report accidental damage to a parked vehicle.

The findings of the survey, begun in 1978 in nine western countries, show that belief in sin is highest in Northern Ireland (91 per cent) and lowest in Denmark (29 per cent). More than twice as many Americans as Europeans believe in hell and the devil.

Even 15 per cent of atheists believe in sin and 4 per cent in the devil.

A preliminary analysis of the findings, to be published in a book in the autumn, is given in the Roman Catholic weekly, *The Tablet*. It shows that 78 per cent of Europeans think there is good and evil in everyone.

The Irish have the most optimistic view of mankind. They think 34 per cent of people are basically good. The figure for the French, who take the most jaundiced view, is 5 per cent.

Most Europeans admit that they sometimes regret having done something they felt was wrong. The Italians and Danes suffer most from such regrets, the French and the Belgians least. The rich regret more than the poor.

The survey, which was carried out by an international team of academics, examines the 'sins' recognised in the West. The Ten Commandments, apart from those about Sunday and worship, are still rated highly.

Killing is top, followed by stealing and honouring parents. Britons rated the prohibitions on adultery and coveting thy neighbour's wife higher than did any other nation.

Most of those questioned cited honesty as the most important quality to be encouraged in children. Only the British put good manners second. For other nations tolerance and respect for other people came next.

The rich are less likely to believe in sin than the poor. The right takes a more cheerful view of the nature of man than the left.

Among parents the strictest attitudes were found among believers in God and regular worshippers. Left-wing parents are less strict than right-wing parents. Parents in lower income groups are tougher disciplinarians than their wealthier counterparts.

Professor Jan Kerkhofs, a Jesuit priest at Louvain University, in Belgium, who is director of the project, said last week that between 1,200 an 2,000 people had been questioned in each country and the findings were still being analysed.

(The Observer: 27/2/83)

Which of the 'sins' mentioned in the text do you think are most serious, and why?

I think that there is good and evil in every one is the most serious sin mentioned in the te,

I think that is the true ~~and~~ because all ca. See some body one day is very nice and very friendly person but one day you mig find him in a very bad hapit.

EXP 121

APPENDIX I

The person himself is some time devil.
He/she kills, steals, says lies, tells sheats
and does many bad things. EXP 121 APPENDIX I
The person himself who ~~mar~~ is making the
guns, bomps, firs, wars, ...

On other hand you find mankind who
is friendly, you find love, family,.
hospitals, sciences, helps,.

Here also the person ~~is the~~ has
a good inside him.

and also ~~we~~ that's right ~~the~~ people
regret if they do some thing wrong.
If some body has done some thing wron
you see him one day is regreting for th
and the rich people do that more tha
poor people because poor people ~~the~~ don'
have ~~every~~ many things like the rich pe
also that is serious and I think it's,

ENGLISH LANGUAGE TESTING SERVICE

Exp 122 APPENDIX I

Band

PHYSICAL SCIENCES MODULE

Test Cent: ..... Date ....

Candidate .....

Answer BOTH questions. You have 40 minutes. Please allow 25 minutes for the first answer, and 15 minutes for the second.

Question 1: New scientific processes often meet with opposition because of the pollution they cause to the environment, an example of which is referred to in Section 2 of your Source Booklet. As a scientist, how would you defend the continual use of such potentially harmful processes? (Write 15 to 20 lines).

122

The phisical facts of pollution can be measured by using scentific equipments, and scientists know the process of the facts. Engineers who know the scientific knowledge only can develop facilities which reduces this harmful processes. On the other hand, politicans and exectives of companies have a force to decide the use of the benefitable but harmful processes. The decision must be or would reflect the will of people who are enjoying and are harmed by the process.

Therefore, the scientists only can give people proper information about the process, and the engineers only can give people proper information about the technology and the cost of preventing the harmful effects.

For me, the question above does not make sence. The choice of continual use of such potentially harmful process or cutting off the use of the process does not depend on the scientis.

Scientis want to know every thing in a rational way. The knowledge obained by this way is so repeatable and testable, or reliable that this knowledge have a power. The way how we use the power is not the responsibility of the scientists.

500